Heterogeneity and Context-Specificity in Biological Systems

Oren Litvin

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2014

ABSTRACT

Heterogeneity and Context-Specificity in Biological Systems

Oren Litvin

High throughput technologies and statistical analyses have transformed the way biological research is performed. These technologies accomplish tasks that were labeled as science fiction only 20 years ago – identifying millions of genetic variations in a genome, a chip that measures expression levels of all genes, quantifying the concentration of dozens of proteins at a single cell resolution. High-throughput genome-wide approaches allowed us, for the first time, to perform unbiased research that doesn't depend on existing knowledge. Thanks to these new technologies, we now have a much better understanding on what goes awry in cancer, what are the genetic predispositions for numerous diseases, and how to select the best available treatment for each patient based on his/her genetic and genomic features.

The emergence of new technologies, however, also introduced many new problems that need to be addressed in order to fully exploit the information within the data. Tasks start with data normalization and artifact identification, continue with how to properly model the data using statistical tools, and end with the suitable ways to translate those statistical results into informative and correct biological insights. A new field – computational biology – was emerged to address those problems and bridge the gap between statistics and biology.

Here I present 3 studies on computational modeling of heterogeneity and context-specificity in biological systems. My work focused on the identification of genomic features that can predict or explain a phenotype. In my studies of both yeast and cancer, I found vast heterogeneity between individuals that hampers the prediction power of many statistical models. I developed novel computational models that account for the heterogeneity and discovered that, in most cases, the relationship between the genomic feature and the phenotype is context-specific – genomic features explain, predict or exert influence on the phenotype in only a subset of cases.

In the first project I studied the landscape of genetic interactions in yeast using gene expression data. I found that roughly 80% of interactions are context-specific, where genetic mutations influence expression levels only in the context of other mutations. In the second project I used gene expression and copy number data to identify drivers of oncogenesis. By using gene expression as a phenotype, and by accounting for context-specificity, I identified two novel copy number drivers that were validated experimentally. In the third project I studied the transcriptional and phenotypic effects of MAPK pathway inhibition in melanoma. I show that most MAPK targets are context-specific – under the control of the pathway only in a subset of cell lines. A computational model I designed to detect context-specific interactions of the MAPK pathway identified the interferon pathway as a major player in the cytotoxic response of MAPK inhibition.

Taken together, my research demonstrates the importance of context-specificity in the analysis of biological systems. Context-specific computational modeling, combined with high-throughput technologies, is a powerful tool for dissecting biological networks.

Table of Content

List of Figures

List of Tables

Acknowledgments

I owe great gratitude to many people for supporting my research, spending significant amount of time training and helping me, and generally being there to support me throughout my PhD.

First and foremost, I am deeply indebted to Dana Pe'er. I joined Dana's lab 7 years ago as a programmer. Having no knowledge in biology and very limited experience with machine learning, Dana took it upon herself to teach and train me, and to transform me into a computational biologist. I can never be thankful enough for all her effort, attention and dedication.

I cannot thank Neal Rosen enough. After several years in a computational lab, I asked Neal to join his lab and receive experimental biology training. Neal agreed, and quickly became my co-mentor. My experience in the Rosen lab has given me the chance to explore aspects of biology I never knew existed. I want to thank Neal for his insights, guidance, harsh criticism, and for teaching me how to doubt everything and everyone.

I owe a debt of gratitude to my friends and colleagues for all their help, support and training they have graciously given me throughout the years. BJ Chen, my dear friend, for his endless help and support. Helen, for always giving the right advise at the right time, and for always being there to listen to my complaints (and there were many). Sarit, my "ifcha-mistabra" – it is always nice to know there is an optimist out there. Special thanks to Christine, Madhavi and Noel, for spending days, weeks and months training me in the wet lab and teaching me everything I know. Felix, Jacob, Elad, Poulikos, Margo and Rona for allowing me to pick their brains whenever I needed something. Tanya and Mark, the best lab technicians, thank you.

I would also like to thank my thesis committee members – Carol Prives, Harmen Bussemaker and Ramon Parsons, as well as the Howard Hughes Medical Institute for supporting my PhD studies. Many thanks to Meehan Crist, for teaching me scientific writing.

I would not have come this far without the support of my dear friend Inbal.

Finally, I'm thankful for the love and support from my close family - my parents, my sister and her family, my brother, and, of course, Ran.

To Ran

לרן

Introduction

Biological systems are of vast complexity, with tens of thousands of molecules participating in a large network to determine the behavior, response and traits of a cell. Pharmacological inhibition of a pathway, changes in cell environment, and even small changes in the concentration of just one protein can trigger a cascade of events that affects the most fundamental cellular features, such as differentiation, metabolism, and cell growth and death. It is therefore not surprising that even closely related systems, such as tumors from the same patient but from different sites, show substantial genomic and phenotypic heterogeneity. My work aims at using genomic features to predict and explain the phenotypic heterogeneity of various biological systems.

Explaining and predicting phenotypes are at the heart of biology research. Phenotypic heterogeneity can be observed across all systems and conditions. For example, some individuals are more susceptible to a specific disease than others; tumors respond differently to treatment, and their metastasis patterns vary between patients; certain cells can adapt to environmental changes, proliferate and differentiate, while other cells are in a terminal fixed state. Understanding the underlying mechanisms of phenotypic heterogeneity is necessary in order to control and influence the phenotypes, develop new treatments, and choose the best treatment for individuals.

Gene expression - predictor of phenotype and phenotype to predict

Many studies focus on the prediction of a phenotype by genotype[1] (figure 1A). These studies aim to identify genetic features - Single Nucleotide Polymorphism (SNP), Copy number variations (CNV) or somatic mutations - that cause or correlate with a phenotypic outcome[2,3]. In other words, is the phenotype a function of a genotype ($P=f(G)$)?

The scoop of these studies is very wide. Genome Wide Association Studies (GWAS) use statistical tests to scan the genome for genetic features, usually SNPs or CNVs, which explain a trait or disease, such as blood pressure[4], schizophrenia[5] and others. In the cancer field, studies

aim to identify germ-line or somatic genetic features that can predict and/or explain tumor progression or response to treatment.

However, the genetic makeup is not always the best predictor of a trait. Additional factors, such as chromatin modifications, inherited cell state (e.g. cell type) and environmental conditions, affect cellular state and can influence the phenotypic outcome[6] (figure 1B). Therefore, genetic features alone typically fail to fully explain the phenotypic variance, and the so-called "missing heritability", the phenotypic variance that is not explained using genetic feature alone, remains the biggest hurdle in genotype to phenotype studies[7-9]. To identify the "missing heritability", and to better predict and explain phenotypic variance, one has to take into account these additional epigenetic factors. However, it is infeasible, and for some features impossible, to measure them for all individuals.

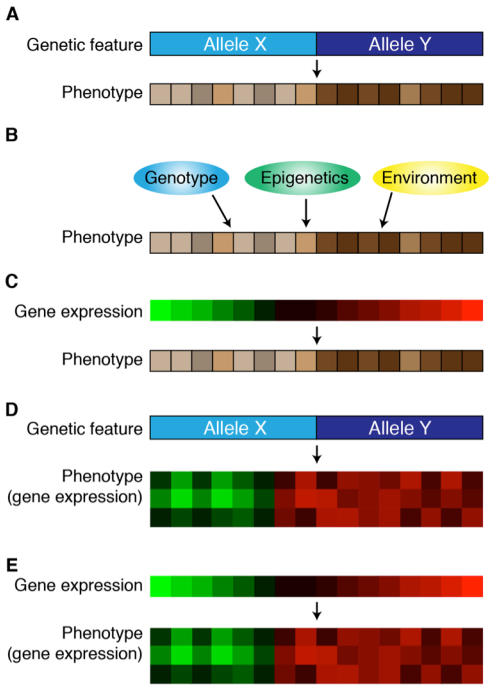


Figure 1 – Gene expression as a phenotype and a predictor of a phenotype
Gene expression integrates the effect of multiple genetic and epigenetic factors, and is a proxy to cellular state. **A.** Classic genotype to phenotype association, where a genetic allele is used to predict phenotype. **B.** Other factors, including epigenetics and environmental cues, also influence the phenotype. **C.** Expression of a gene is used to predict a phenotype. **D.** Gene expression cluster serves as a phenotype to be predicted. **E.** Gene expression is used both as a predictor and as a phenotype - expression of one gene predicts the expression of others.

Another approach, instead of measuring each and every epigenetic and environmental feature, is using gene expression as a proxy to them. Epigenetic factors, together with the genetic background, influence the expression of many genes[10,11]. Therefore, gene expression can be viewed as an "integrator" of several factors, and can therefore "represent" them. Taken together, gene expression can be used as a predictor of phenotypes in association studies[12] (figure 1C).

The use of gene expression has several advantages. *First*, high-throughput technologies to measure gene expression levels, such as microarrays or RNA-sequencing, are relatively cheap, and provide information on tens of thousands of transcripts at once. *Second*, expression of genes behaves in concert, and genes are highly correlated with each other. By clustering genes based on their expression, noise of both the biological system and the measurement decreases, while statistical power increases. *Third,* used as a proxy, gene expression significantly shrinks the feature space from millions of SNPs and complex epigenetic profiles to just ~30000 gene expression patterns, drastically relieving statistical burden. *Fourth*, gene expression can be linked to protein activity, which can be critical for the mechanistic understanding of a response[13], but is much harder to measure with high-throughput methods.

While expression patterns can be used to predict phenotypes, they can also be considered as the phenotype itself. Expression levels reflect cellular states and can inform us on the activity of various pathways and cellular functions that we wish to predict and explain[13]. For example, oncogenic activation of a pathway alters the expression of its downstream genes[14]. We can use expression patterns as cellular phenotypes, and associated them with the genetic background[15] (figure 1D). Genetic features that influence expression levels are also called eQTLs (expression quantitative expression loci)[16].

Finally, as gene expression can be used as phenotypes themselves and predictors of phenotype, one can design a study that uses gene expression for both (figure 1E). However, such studies have several drawbacks. Specifically, the high correlation between gene expression profiles makes it difficult to identify the best predictive feature, and unlike DNA features, gene expression can be both the cause and the effect. By addressing these pitfalls via the integration of DNA based data, expression-to-expression association studies, such as CONEXIC (described in chapter III), provide an important framework for identifying both important features and unknown phenotypes.

Taken together, gene expression is a powerful resource that can be used in many applications of biological research. In my work, I use gene expression patterns as both the phenotypes and

the predictors of phenotypes. My results demonstrate that gene expression is a versatile feature - it is a proxy to the metabolic state of the cells, it can be used to identify driver oncogenes, and in some case it is associated with resistance and sensitivity to targeted therapy.

Context-specificity of biological systems

Genotype-to-Phenotype association studies aim at identifying a genetic (or genomic) feature that predicts the phenotype. In other words, the goal is to represent a phenotype P as a function of a genomic feature G: $P \cong f(G)$.

In most cases, however, multiple features influence the phenotype[17]: $P \cong f(G_1, G_2, \ldots, G_n)$. To estimate the influence of multiple features, most association studies use additive models, in which the phenotype is a linear combination of the features: $P \cong aG_1 + bG_2$[18]. The underlying assumption of additive models is that all features exert influence in all individuals.

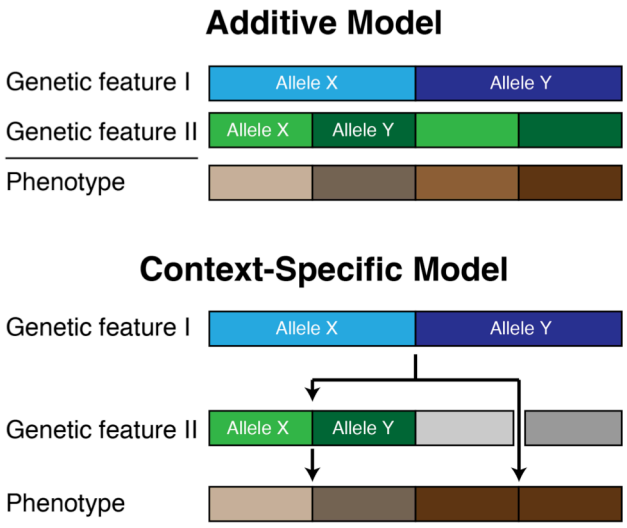


Figure 2 – interactions between genetic features
In an additive model, the phenotype is a linear combination of two genetic features – $P = aG_1 + bG_2$. The effect of each feature is similar, regardless of the context of the other feature. In the context-specific model the second feature only affects when the first feature is of a specific allele. When the first feature is of the other allele, the second feature bares no influence (grayed out), and the phenotype depends only on the first feature.

Using examples from three biological systems, my results demonstrate that in many cases the relationship between the genomic features is ***context-specific.*** In these cases, the predictive model can be viewed as a tree (figure 2). First, the value of the main feature is examined, and only in **specific contexts**, i.e. the individual has a specific allele of a gene, the second genetic feature exerts an influence.

The concept of context-specificity is by no means novel in biological research. For example, PTEN mutations only arise in the context of BRAF mutations in melanoma[19]. However, my work

demonstrates how prevalent these types of interactions are. For example, by using context-specific models, I was able to identify 5 times more interacting eQTLs than with additive models (chapter II).

Importantly, these tree-based models can't be mathematically represented by an additive function. Moreover, such relationships add another layer of complexity to the model – not only that we need to identify the most predictive features, we also need to identify the contexts in which they exert influence. In this thesis I present three computational models, GOPLH, CONEXIC and COSPER, each taking a different approaches to identify the contexts and the features. By applying my novel methods to data of several types and sources, I was able to identify key genomic features, such as driver genes in cancer and genetic mutations that are associated with resistance to treatment.

Context-specific eQTLs

In chapter II, I explore the landscape of genetic interactions in S. *cerevisiae* using gene expression as a phenotype. My results show that roughly 80% of genetic interactions are context-specific.

Yeast is used as a model organism to better understand association between genotype and phenotype. The readily available genetic and genomic data of a panel of related individuals (strains) provides a testing ground for hypotheses and the design of mathematical models for linkage and association studies. Here, I use genetic loci and associate them with gene expression clusters to identify eQTLs.

Previous work on eQTLs in this data searched for one locus-one gene pairs. However, it was demonstrated that one locus does not fully explain the expression variance of a gene. Therefore, Brem et al.[17] expanded their search for multiple loci (two in their study) using the additive model ($P \cong aG_1 + bG_2$, figure 2). However, this model could only explain a fraction of the variance in gene expression.

In search of the missing heritability, I developed GOLPH, a non-additive statistical model, to characterize the landscape of interacting eQTLs. GOLPH introduces two novel concepts that greatly improve prediction:

1. Instead of associating only one gene as a time, GOLPH uses clusters of co-expressed genes. This approach greatly reduces the statistical burden of the model, while also reducing biological and measurement noise.
2. To examine the effect of multiple loci, GOLPH uses a tree-based model, which includes both additive and context-specific interactions (figure 2).

Using GOLPH, I was able to identify 5 times more interactions than previously published methods, mostly due to the incorporation of context-specific models. I later expanded GOLPH, in collaboration with Anat Kremer and Itsik Pe'er, to work on human data[20]. These results demonstrate the prevalence of context-specific interactions across biological systems.

Identifying context-specific drivers in melanoma

In chapter III I present a work on driver mutations in melanoma. It serves as a logical extension of my finding in chapter I, while introducing new concepts specific to cancer biology. The work, which was done jointly with Dr. Akavia, aims at identifying DNA copy number drivers in cancer.

During tumorigenesis, due to genomic instability, the DNA accumulates dozens of aberrations, including point mutations, amplifications, deletions, translocations and others. Each aberration can lead to activation of oncogenes or repression of tumor suppressors in different ways. Point mutations, for example, can lead to a loss-of-function of a tumor suppressor or gain-of-function of an oncogene[21]. DNA amplifications lead to overexpression of the genes in the amplified region, which in turn can over-activate an oncogene[22]. However, only a small percentage of those aberrations, i.e. driver mutations, contribute to the fitness of the tumor. Other mutations are fixed in the tumor's DNA but don't provide any fitness advantage (i.e. passenger mutations).

Identification of driver aberrations, i.e. distinguishing them from passengers, is essential for the development of new treatments and treatment regimens.

Large projects, such as The Cancer Genome Atlas, use high-throughput technologies to quickly identify all DNA aberrations in large panel of tumors. However, as each tumor harbors dozens to hundreds of aberrations[23,24], distinguishing drivers from passengers requires additional computation, data or knowledge and remains a major hurdle in putting these large datasets in use.

Typically, identification of drivers is based on their frequencies in a panel of tumors. Since the likelihood of a specific locus to mutate during tumorigenesis is very low, an aberration that recurs in many tumors is likely to be a driver mutation. Methods based on aberration frequencies, such as GISTIC[25], have proven to be successful and identified dozens of oncogenes and tumor suppressors[26,27]. However, not all driver mutations are overrepresented in the tumor panel, and approaches based solely on frequencies tend miss them.

Frequency based method fail for two reasons, context specificity and low penetrance. Certain genes are drivers only in a specific context. For example, PTEN is only deleted in BRAF-mutated melanomas. Therefore, PTEN might be deleted in 60% of BRAF tumors, but its overall deletion frequency in melanoma is only 30%. By assessing the significance threshold for all tumors combined, PTEN might fall below threshold and be classified as a passenger. In order to define true significance thresholds of driver aberrations, one must first define the **context** in which the driver is essential. The problem of low penetrance arises for a different reason – certain aberrations drive a malignant phenotype, but do so in a small subset of tumors. To identify such driver, one can identify a malignant phenotype that exists in a subset of tumors, and associate genetic aberrations with it.

Copy number aberrations present an additional hurdle. Unlike point mutations that are specific for a gene, copy number aberrations affect dozens of genes at once, amplifying or deleting large genomic regions. Therefore, even after identifying “driver” aberrant regions, one still has to identify the driver gene within a large region (on average, a copy number aberrant region in melanoma contains 23 genes).

In our work, we aimed at identifying copy number based drivers. We designed a computational method - CONEXIC - that attempts to overcome the three pitfalls of frequency-based driver identification approaches – low penetrance, context-specificity and large aberrant regions. In short, CONEXIC starts with a very permissive frequency threshold and tests many putative genes as drivers. Then, to assess the contribution of each putative driver to a malignant phenotype, we associate the gene expression pattern of the gene with a malignant phenotype defined by a gene expression cluster. Additionally, to account for context-specific drivers, we define contexts based on gene expression patterns, and test each putative driver in each of the contexts.

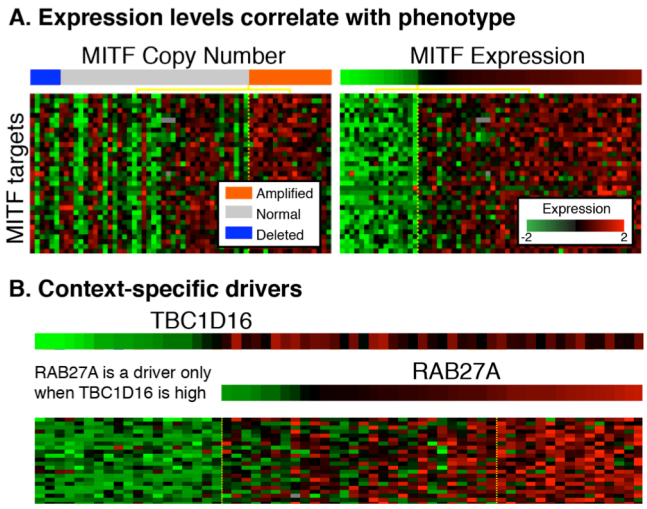


Figure 3 – Identifying CNA drivers in melanoma
Our computational model - CONEXIC – identifies copy number drivers. **A.** CONEXIC identifies the driver genes within aberrant regions using their gene expression profile. Here a known oncogene, MITF, and its targets are shown. The expression of the oncogene, and not its copy number profile, predicts the expression levels of its targets. **B.** CONEXIC also identifies the context in which a driver exerts an influence. In this case, RAB27A only influence its putative targets when TBC1D16 is highly expressed.

CONEXIC was the first method to combine gene expression patterns and genetic information to distinguish drivers from passengers. Additionally, CONEXIC includes three major advancements over the current paradigms and methods in the field (figure 3):

1. The gene expression of a gene, and not its copy number profile, is used to assess its contribution to the malignant phenotype. Figure 3A demonstrates that the expression of MITF, and not its copy number profile, is correlated with the expression of MITF's known targets. If a gene is important for tumorigenesis, cells will find various ways to over-activate it, and copy number is just one of those ways. By using expression and not copy number, CONEXIC integrates the effects of various aberrations that lead to over-activation of the driver.

2. The malignant phenotypes are represented by gene expression clusters, which are used to evaluate the contribution of the driver genes. By identifying putative drivers whose gene expression correlate with the phenotypes, CONEXIC can distinguish between drivers and passenger that share the same aberration profile.
3. The effect of driver aberration is assessed in the context of other drivers, under the assumption that not all drivers are important in all tumors. This leads to the identification of context-specific interactions between drivers.

Using CONEXIC we identified two novel drivers, RAB27A and TBC1D16, both involved in protein trafficking, and experimentally validated tumors' context-specific dependency of them.

Context-specific interactions of the MAPK pathway in melanoma

In the fourth chapter I present my work on phenotypic heterogeneity of drug response in melanoma. While CONEXIC aimed at identifying drivers, here I investigated how genetic aberrations alter the response of melanoma to inhibition of a key oncogenic signaling pathway.

The RAS-dependent extracellular signal-regulated kinase (ERK1/2) Mitogen Activated Protein Kinase (MAPK) pathway transmits signals from growth factor receptors to the nucleus, and plays crucial role in various cellular functions such as proliferation, apoptosis, differentiation and more. In melanoma, the pathway is constitutively active by mutations in 70-90% of tumors. Inhibition of the pathway leads to a dramatic decrease in proliferation and an increase in cell death, both *in vitro* and in patients. Recently, several highly potent and specific drugs that inhibit the pathway have been approved for clinical use [28,29]. However, not all patients respond to the drugs, and the responses in patients that do show some response are heterogeneous[30]. The molecular mechanisms that underlie this phenotypic heterogeneity are still not well understood. In this work I investigated context-specific interactions of the MAPK pathway in order to pinpoint the molecular mechanisms underlying phenotypic heterogeneity.

The MAPK pathway in melanoma

The ERK-MAPK signaling pathway transmits signals through a cascade of kinases – RAF->MEK1/2->ERK1/2. RAS, a small GTPase, is required for the physiological activation of the kinase cascade and is considered a member of the pathway (figure 4A). Upon activation of several growth receptors, the pathway, and most importantly ERK1/2, undergoes a rapid and strong burst of activation that triggers cell division and suppresses cell death[31].

The MAPK pathway is deregulated by various molecular mechanisms in about a third of all human cancers, including breast, pancreas, lung, melanoma and others[32]. In melanoma, the pathway is constitutively active in 70-90% of tumors, by point mutations in NRAS (~20%) or BRAF (~50%), or by NF1 loss (~20%), although the latter has yet to be directly implicated in MAPK-dependent melanogenesis[33] (figure 4B).

The list of reported direct and indirect ERK targets is very long and involves almost every cellular pathway and process, but varies significantly depending on the context and conditions tested[34-37]. To further complicate things, ERK acts both as a kinase that directly alters activation of its targets, and as a transcription co-factor to induce or repress transcription of genes[38]. Deciphering the complex network of MAPK targets and the contexts in which they are regulated is necessary for the development of new drugs and treatments for melanoma patients.

Phenotypic heterogeneity in response to MAPK pathway inhibition

In recent years, several highly specific and potent small molecule inhibitors of BRAF, MEK and ERK were developed[39-41]. This extended panel of drugs significantly enhanced our ability to investigate the MAPK pathway, its downstream targets, and the phenotypic consequences of its inhibition.

The first potent and specific drugs to be developed were MEK inhibitors, and therefore most of the published literature examines the effects of MEK inhibition. *In vitro* studies using melanoma cell lines show a dramatic growth inhibition following MEK inhibition in almost all MAPK-activated melanomas[42]. However, cytotoxic (i.e. cell death) response varies significantly between cell lines and conditions, some cell lines show almost no cytotoxic response, while others undergo massive cell death[42]. The mechanisms by which MAPK activity suppresses apoptosis, and the molecular consequences of MAPK inhibition that lead to cell death are still not fully understood and contradictory explanations exist[42-44].

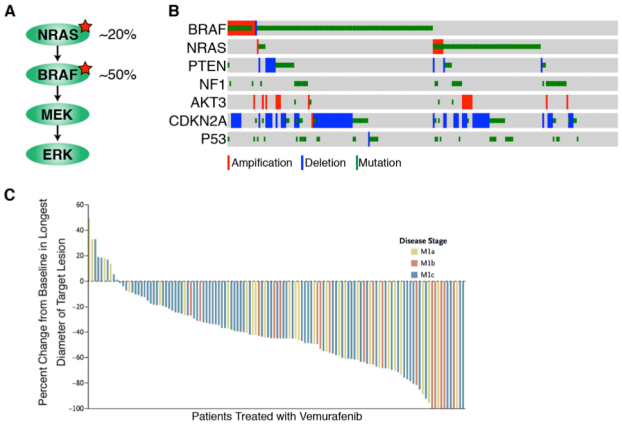


Figure 4 – MAPK pathway in melanoma
A. The ERK-MAPK pathway transmits signals from growth receptors via a cascade of kinases. Numbers show approx. percentage of mutations in melanoma tumors. **B.** Mutation and copy number data from the TCGA project (source: bioportal, MSKCC). Each column represents one tumor. C. Clinical results of the Vemurafenib clinical trial. Each bar represents the maximal volume change of one lesion in one patient. Source: Sosman et al. NEJM 2012

Clinical results also demonstrate response heterogeneity. The drug that attracted the most interest is vemurafenib[30] (PLX4720), a highly specific molecule that targets the BRAF-V600 mutation that exists in ~50% of melanoma patients. Compared with other available melanoma treatments, the drug showed remarkable clinical results and was quickly approved for clinical use, but not all BRAF-mutant patients respond to the drug[30] (figure 4C). However, the reasons for clinical heterogeneity can be far more complex than the *in vitro* models, and can include interactions with the immune system, genetically heterogeneous tumors, drug delivery and more.

Context-specific interactions of MAPK in melanoma

My project addressed the response heterogeneity of MAPK-activated melanoma to MEK inhibition in vitro. By studying the underlying mechanisms of cytotoxic and cytostatic responses, I

hoped to identify some of the reasons for the varied clinical response, and suggest new targets and treatments that will improve clinical outcome.

Many computational-based studies have attempted to identify genetic and genomic features that can explain and predict response to drugs[45,46]. These association studies use high-throughput data, such as gene expression and mutations, and correlate them with a quantitative phenotype, such as growth rate. Thus far, this approach has failed to identify good predictors of cytotoxic and cytostatic responses for several reasons. First, growth-based phenotype is a result of several independent cellular phenotypes (growth arrest, death), and models aiming to predict a response ought to predict each cellular response independently of others. Furthermore, these studies use mathematical models that assume that the same molecular mechanism underlies a cellular response in all individuals, typically even across different cancers, which is unlikely to be the case (see Discussion for a full comparison of the methods).

I took a different approach. I hypothesized that context-specific targets and interactions of the MAPK pathway underlie the heterogeneous responses to MAPK inhibition. Therefore, a better, context-specific, characterization of the network regulated by the MAPK pathway will help identify the reasons underlying phenotypic heterogeneity.

To allow better inference of the network, I used both pre- and post- inhibition data. Most association studies only use steady state expression data and rely solely on correlation to identify MAPK targets. I assumed that supplementing the data with post-inhibition expression profiles would greatly enhance my ability to identify targets. Not only can perturbations change cell state and expose hidden interactions[47], by perturbing the network I am able to infer causality and directionality of the interactions. Therefore, I measured changes in gene expression following MEK inhibition in a panel of MAPK-activated melanoma cell lines, since MEK inhibition fully inhibits the MAPK pathway both in NRAS- and BRAF-mutated cells.

Notably, I found that although all cell lines harbor a MAPK activating mutations, most targets of the pathway are context specific – under the control of the pathway in only a subset of cell lines. To identify these context specific targets, and to assess their contribution to the phenotypic

variance, I developed computational tools that analyze pre- and post- perturbation gene expression data.

Using the computational methods, I identified that the interferon pathway has a key role in the cytotoxic response to MEK inhibition. The computational analysis found that the interferon pathway is either on or off in different cell lines, and identified an interaction between the interferon and the MAPK pathways. An experimental validation showed a synergy between two unrelated drugs for melanoma – Type-I Interferon (IFNα/β) and MEK inhibitor. Moreover, copy number analysis identified that a deletion of the interferon genes predicts sensitivity to MEK inhibition. Taken together, my results showed that the interferon pathway plays an important role in melanoma.

Summary

Context-specificity proved to be the key feature in all my projects. While the biological systems and the questions asked in each project were different, models that account for heterogeneity of the underlying molecular networks were essential to understand phenotypic variance. Analysis of each of the cases above found vast differences in the interactions and dependencies of genes and proteins between individuals (cell lines, yeast strains). Since our goal is to explain and understand human disease in different individuals, such differences must be identified and understood. I believe that a better understanding of context-specific interactions and heterogeneity would enable clinicians to tailor new and unexpected drug combinations to individual patients based on their genetic profile, which may lead to better clinical outcomes.

Chapter I – Review: Genetic and genomics of melanoma

Melanoma – an aggressive form of skin cancer

Metastatic melanoma is an incurable disease, with only 14% of patients with survive for five years[48]. Roughly 80000 people are diagnosed with melanoma annually, of which 10000 are metastatic.

Melanocytes, the pre-cursors of melanoma, are derived from the neural-crest cell layer. During tumorigenesis they develop and form nevi, and the subsequent development of dysplasia, hyperplasia, invasion, and metastasis[49].

Cancer can be seen as a Darwinian evolutionary process, with genetic and epigenetic changes providing cells a fitness advantage to proliferate, survive and metastasize. All cancers, including melanoma, share a common pathogenesis. A premalignant cell gradually transforms, through the acquisition of oncogenic aberrations, into a malignant tumor with the ability to metastize[50] . In melanoma, UV radiation is the major cause for the accumulation of DNA mutations[51].

Recent studies portrayed the landscape of genetic alterations in melanoma and helped identify several key oncogenes and tumor suppressors[52]. Of main interest are mutations in NRAS and BRAF that activate the MAPK pathway in more than 70% of tumors. Drugs that target the pathway have been recently approved for clinical use and show exciting clinical results. However, clinical results vary significantly between patients, and our limited understanding of melanoma's molecular biology hampers our ability to design new drugs and treatments.

In this chapter I review our current understanding of molecular biology of cancer in general and of melanoma specifically. I start with the description of the process that drives cancer – mutagenesis, and summarizes the molecular events underlying it. Then I discuss how mutagenesis, by altering the activity of "driver" genes, leads to tumorigenesis. I then review the different computational and biological methods to identify these keys drivers. Finally, I provide a brief review of melanoma drivers.

DNA integrity and mutagenesis

Cancer is a rapid evolutionary process, driven by accumulation of mutations and aberrations in DNA. Those aberrations are caused by several mechanisms, including error-prone DNA synthesis, external carcinogens such as tobacco and UV radiation, internal carcinogens such as reactive oxygen species, loss of DNA integrity checkpoints and others.

Mutations that contribute to proliferation and fitness of a cell are selected during the evolutionary process. These mutations support numerous processes, including proliferation, overcoming cell death, angiogenesis, rapid metabolism and more[53]. While we still do not have the tools to identify all key aberrations, we do have the molecular understanding of how certain aberrations support ontogenesis.

Type of aberrations

Genetic instabilities and lack of genetic integrity, together with the effects of carcinogens, lead to the accumulation of genetic aberrations. There are several categories of genetic aberrations, and while their mechanisms of action is different, they all lead to the activation of oncogenes or repression of tumor suppressors:

Point mutations – single DNA base mutations account for most DNA aberrations in cancer. The most prevalent aberration is base substitution, which account for 95% of these mutations[54]. The others, accounting for 5% of point mutations, are single nucleotide insertion and deletion. In melanoma, initial reports based on whole-genome sequencing identified roughly 10000-80000 mutations per cancer genome[51,52].

While most point mutations are in "gene-deserts" and have no phenotypic effects, others can lead to several molecular effects:

- Oncogene activation – the first point mutation that was shown to have oncogenic activity was identified in the gene HRAS. The mutation, a substitution of G>T in codon 12 of the gene[55], leads to constitutive activation the MAPK pathway. Since then, additional point mutations that lead to constitutive activation of oncogenic proteins were identified, such as BRAF[56] and PIKC3A[57].

- Loss of function – point mutations can lead to nonsense mutations or amino acid substitution that inhibit the expression or activity of a gene. An example is TP53, a tumor suppressor mutated in ~40% of human cancers[58].
- Splicing – point mutations in splicing sites can lead to both loss- and gain- of function in proteins[59,60]. However, since our understanding of splicing mechanisms is still limited, it is hard to identify such mutations and understand their molecular consequences.

Copy number aberrations – tumors often display gain or loss of chromosomal regions[61]. In some cases, entire chromosomes are lost or gained (aneuploidy[62]), for example, gain of chromosome 8 in 20% of acute myeloid leukemia[63]. Other cases show a focal deletion or focal amplification of a chromosomal region or a specific gene, for example, amplification of HER2 in breast cancer[64]. Copy number aberrations (CNA) lead to under- and over- expression of genes, and alter their activity levels.

Translocations and fusion genes – chromosomal translocation is caused by fusion of subparts of chromosomes. While the definition usually refers to a fusion of nonhomologous chromosomes, for example, fusion of chromosomes 14 and 8 in Burkitt's lymphoma[65], chromosomal deletions that lead to a fusion of two genes on the same chromosomes are also characterized as translocations[66]. Oncogenic translocations usually fuse a promoter of a gene that is highly expressed in the cancer's cell lineage to an oncogene. Increased expression of the oncogene leads to constitutive activation of its pathway. The first genetic aberration that was ever identified in cancer was the "Philadelphia chromosome" - a translocation that fuses BCR to ABL in chronic myelogenous leukemia[67].

Short insertions and deletions (indels) - insertion or deletion of a small number of nucleotides. If the resulting change in number of nucleotides is not a multiple of 3, it will results in a frameshift mutation that is likely to cause an early stop codon and a truncated protein. As an example, TGFβ receptor is inactivated in colon cancer primarily due to short indels[68].

DNA instability crisis

It has been argued that a global event of "genomic instability", which leads to a rapid accumulation of mutations, is required for the generation of sufficient number of oncogenic mutations[69,70]. However, others have disputed this claim, demonstrating that some tumors have a normal rate of mutagenesis[71].

Global and catastrophic genetic instability can arise by several mechanisms. Hereditary defects in genes controlling the mismatch repair mechanism, such as MLH1 and MSH2, lead to rapid accumulation of point mutations in colon cancer[72,73]. Defects in the identification and repair of double-stranded DNA breaks, commonly caused by hereditary mutations in BRCA1 and BRCA2, also lead genetic instability and account for 5-10% of breast cancer[74,75]. A chromosomal "crisis", caused by shortening of telomeres of aging cells, is another major source of genetic instability and can lead to a rapid accumulation of drastic aberrations of the DNA[76].

Molecular basis of genetic alterations

The molecular basis of genetic aberrations in cancer can be roughly divided into 3 categories: 1. Sporadic and random aberrations caused by imperfections in the synthesis and mitosis cellular mechanisms; 2. Carcinogen-induced mutations, such tobacco and UV; 3. Systematic mutation accumulation due to catastrophic failures of the mechanism responsible for genomic integrity.

The second and third categories are of the most interest. A better understanding of carcinogen-induced mutations will help to direct prevention guidelines, such as sunscreen to reduce exposure to UV radiation. For example, recent studies show a high rate of TC>TG mutations in breast cancer, but the carcinogen or molecular mechanism responsible for these mutations hasn't been found[77].

Errors in mechanisms that maintain genetic integrity are also of interest. First, tumors acquire resistance to treatment by further accumulation of mutations. Identifying the molecular mechanisms of systematic mutagenesis will allow development of drugs that will prevent or slow

down mutation rate. Second, familial cancers are often caused by heredity mutations in genes involved in genetic integrity, such as MSH2, a gene involved in the mismatch repair mechanism. Pinpointing such heredity mutations will allow identifying the population at risk and will guide early detection and prevention efforts.

While some mutagenic mechanisms are well described, such as UV-induced mutations, others, such as focal amplifications, are not understood as well. Here I give a brief overview of the different molecular mechanisms underlying mutagenesis:

UV- and carcinogen-induced point mutations – Carcinogens and ultra-violet radiation lead to DNA damage by specific molecular changes of nucleotides. Ultra-violet, for example, leads to two types of lesions - cyclobutane pyrimidine dimers and 6-4 photoproducts[78]. Acrolein, a carcinogen found in cigarettes, irreversibly binds to DNA and interferes with new strand synthesis[79]. These DNA aberrations are repairable. Two DNA repair mechanisms - nucleotide excision repair (NER) and base excision repair (BER) – are responsible to the identification and repair of these lesions[80,81]. However, when these repair mechanisms fail to identify and/or repair the lesions, the mutations persist and interfere both with the synthesis of a new strand and with RNA expression.

Different organs are exposed to different carcinogens, which influences the type of mutations in tumors of those cell linages. For example, 60% of point mutations in melanoma are C>T, a result of failed repairs of UV-induced lesions[51].

Replication-based point mutations and Microsatellite instability – DNA synthesis is an error prone mechanism. Although the process is characterized by high fidelity, most mismatch base errors repaired during synthesis, some bypass the multiple mismatch repair mechanisms and are fixed in the nascent strand[82,83]. However, events that alter the activity of the repair mechanisms, such as downregulation or mutations in proteins that form complexes that identify mismatches, can significantly increase synthesis mutation rate. For example, 15% of sporadic colon cancers display downregulation of genes in the MSH family, which leads to accumulation of

small indels in highly repetitive sequences, also known as microsatellite instability[84]. Importantly, patients with microsatellite instability show distinct clinical responses[85].

Aneuploidy – Gain and loss of whole chromosomes occur in all human cancers[86]. It is possible that the rate of aneuploidy is higher than currently estimated , as techniques that estimate aneuploidy miss loss-of-heterozygosity and whole chromosome duplication events, which are prevalent in some cancer types[87]. The molecular mechanisms underlying aneuploidy are not well understood, but it is hypothesized that they arise due to improper monitoring of cell cycle progression and errors in centrosome/microtubule formation[62]. Mutations in mitotic checkpoint genes have also been identified[88].

Focal copy number aberrations – although somatic focal copy number aberrations are frequent in cancer[61], the underlying molecular processes that lead to these aberrations are not well characterized[89]. It is currently thought that double-stranded breaks, together with the lack of robust genomic checkpoints, lead to focal DNA amplification[90].

From mutations to oncogenes

In the early days of cancer research, before cloning, sequencing and high-throughput technologies, epidemiologic studies of carcinogenesis suggested that mutations in the genome are responsible, involved or essential for the initiation and progression of cancer[91], but direct evidence was missing. In his seminal study on familial retinoblastoma, Knudson has laid down the statistical framework that demonstrated the "two-hit" hypothesis, suggesting that tumors arise after the loss of two copies of a "tumor-supprosser"[92]. The genomic location of the retinoblastoma gene, Rb1, was only identified 7 years later, using karyotyping techniques on just two patients[93].

Many other tumor suppressors and oncogenes were discovered prior to the use of high throughput technologies. In 1973 Rowley has identified a chromosomal aberration in 9 leukemia tumors, which is now known as the Philadelphia chromosome[67] using karyotyping. The discovery of oncogenic viruses and the research on the molecular biology underlying virus-induced

tumorigenesis demonstrated that the human genome harbors several pre-oncogenes. Later, by comparing viral oncogenes with human pre-oncogenes, the first single nucleotide oncogenic mutation was characterized in the gene HRAS[94]. Linkage analysis was also used to identify oncogenes and tumor suppressors in familiar cases of cancer. For example, studies on the autosomal recessive disease Xeroderma pigmentosum identified 7 DNA repair genes responsible for its manifestation, correspond to 7 disease groups[95].

Prior to high-throughput technologies, identifying oncogenes and tumor suppressors involved linkage analysis, molecular biology studies and a great deal of luck. In the era of high-throughput technologies, it is possible to identify all genetic and epigenetic alterations in hundreds of tumors. However, while these technologies solve one problem – identifying mutations, they create a new one – identifying which of the tens of thousands of mutations each tumor harbors are oncogenic.

Drivers and passengers

Cancer is a result of somatic genetic and epigenetic aberrations that provide the cell with the oncogenic phenotypes required for proliferation and survival. Although the exact number of mutations required for tumorigenesis is still disputed, the numbers range between 4 and 30[91,96,97]. Mutagenesis is a random process, and pre-malignant cells acquire many mutations until genes that support transformation are hit. Moreover, since tumors often lack mechanisms that maintain DNA integrity, they continue to acquire additional mutations[98].

We can therefore classify mutations into two categories – “drivers” and “passengers”. Drivers are genes and mutations that support the tumor, while passengers are mutations that were “fixed” in the genome during tumorigenesis due to random mutagenesis.

The definition of drivers, however, is very complex and needs some refinement:

- Activity: many genes are required for tumor survival and proliferation, and most of them are needed for the physiologic operation of non-malignant cells as well. Therefore, the term “drivers” usually refers to genes with altered activity in tumors – mutated, over-activated, under-expressed, etc.

- Timing: while some altered genes are required for transformation from pre-malignant to malignant cells, they might not be required for the survival of the tumor. Such genes can also be classified as drivers, although they have no functional activity in tumors.
- Function: the range of functions oncogenes support is also very broad. While most studies identify drivers required for proliferation, mutations also support additional hallmarks of cancer, such as evading the immune system and inducing angiogenesis[53].
- Resistance: another class of "drivers" includes genes that play a role in drug resistance. While these genes do not support tumor survival in physiological conditions, they are required for survival under specific drug treatments, and are often acquired only after treatment.

Due to the diverse functionality of drivers, it is hard to define one method that identifies all drivers in all contexts and tumor types. Therefore, there are several different methods to identify drivers, each of which define and identify drivers of different type.

Methods to identify drivers

When sequencing first became available, studies used prior information and focused on likely oncogenes. For example, in 2004 a study that specifically sequenced genes in the PI3K family identified high frequency of activating point mutations in PIK3CA[99]. Another study focused on BRAF, the gene downstream of the then known RAS oncogenes, and identified the V600E mutation in various cancer types[21], and was later shown to activate the MAPK pathway in 50% of melanoma tumors[100].

When studies switched from targeted sequencing to whole-genome technologies, like microarrays and DNA-sequencing, researchers have learnt that each tumor harbors thousands of aberrations. The focus, therefore, had to shift from identifying mutations to identifying driver mutations[101]. New experimental and statistical methods were developed to tackle the new problems introduced by high-throughout whole genome technologies.

Frequency-based methods

Most driver identification methods assume that mutagenesis is a random process, and the likelihood of a specific gene to be mutated across many tumors is fairly low. However if a gene is a driver that is required for tumor survival, mutations in it will be repeatedly selected in many tumors. Therefore if a mutation is observed across several tumors, and the frequency of the aberration is more than expected by chance, the mutation is likely to be a driver that was selected for during the evolution of the tumor.

One of the first and most surprising drivers identified by high-throughput sequencing was IDH1. Mutations in isocitrate dehydrogenase 1 (IDH1) were identified in 5 out of 22 glioblastoma tumors that underwent whole exome sequencing[102]. IDH1 participates in the oxidative phosphorylation cycle, and was the first metabolic driver to be identified. In an extended panel of glioblastoma tumors, IDH1 mutations were found in 12% of tumors, and its mutation predicts better overall survival[102]. Notably, since the gene and its family were never implicated in cancer, only when studies switched from targeted sequencing to unbiased whole genome approaches were the mutations were identified.

Whole genome non-sequencing approaches were also found effective in driver identification. Garraway et al. used high-density SNP arrays to identify copy number aberrations in a panel of melanoma tumors[103]. They identified a focal amplification in chromosome 3, and together with gene expression levels, that pinpointed amplification and overexpression of MITF, a melanocyte-specific transcription factor, which is now classified as a melanoma driver.

Copy number aberrations, however, present a new statistical hurdle. While other types of mutations are specific to a small genomic location, copy number aberrations span genomic regions of various lengths. Some aberrations are focal, confined to an exon or a gene, while others include dozens to hundreds of genes. Therefore, statistical models that assess the significance of the aberration must define the aberrant region, test the likelihood that the aberration is not due to random chance and identify the target driver gene within that region. Several statistical frameworks were developed to address this problem, but most of them solve just half of the problem[25,104]. They identify genomic regions with high recurrence of amplification

or deletion, but those regions often contain dozens of genes. An analysis that integrates multiple data types is required to pinpoint the driver gene within those regions.

The understanding that multiple data types are required for a comprehensive analysis of cancer genomics has led to the initiation of large-scale cancer genomic projects. Spearheading the effort is "The Cancer Genome Atlas" (TCGA) project[24,105]. The project consortium collects hundreds of tumor samples of each cancer type and measures their gene expression levels, copy number aberrations, mutations, translocation, epigenetic state and more. These large data sets give an extensive overview of the genetic landscape of cancer[106]. However, it is clear that these vast datasets will require the development of advanced bio-statistical methods that integrate different data types for the identification of driver mutations, pathway activation and other cellular aberrations that contribute to tumorigenesis.

Association-based methods

Another class to driver identification methods focuses on aberrations that contribute to a specific and known phenotype. The phenotype usually reflects clinically important characteristics, such as overall survival, response to treatment, and metastasis patterns[107-111].

In most studies, the outcome is overall survival and the predictors are either copy number aberrant regions or gene expression profiles. For example, Xie et al. found that gain of chromosome 20q predicts better overall survival in colorectal cancer[110]. Hicks et al. identified a gene expression profile comprised of 95 genes that predict response to tamoxifen in ER+ breast cancer[112].

Association based methods are also used in *in vitro* settings to identify mutations and other aberrations associated with response to treatment. A group led by Andy Minn at the University of Pennsylvania found an interferon-based gene expression signature that predicts response of breast cancer cell lines to radiation[113].

Large-scale association based efforts to identify genetic and genomic determinants of drug response *in vitro* are underway[45,46]. In these studies the growth response to dozens of drugs in hundreds of cancer cell lines is assessed, and then associated with whole genome features, such

as mutations, copy number aberrations and gene expression profiles. So far, however, these studies failed to provide fresh insights into the resistance mechanisms to targeted therapy.

Functional assays

Another approach to identify drivers that contribute to fitness of cancer cells is based on unbiased functional screens using siRNA or shRNA[114-116]. In these assays, a large panel of small interfering RNAs is used to identify genes that are required for proliferation. These assays are performed either under normal growth conditions or under drug treatment. In the latter case, the results help identifying genes that confer resistance to treatment, and can therefore be used to suggest combinatorial treatments. In many cases, in order to gain statistical and biological power, a panel of cell lines is used.

Prahallad et al. used such method to show that EGFR signaling confers resistance to MAPK pathway inhibition in colon cancer[114]. Cheung et al. started the Achilles project, a systemtic effort to identify genetic dependencies across cancer cell lines. Their initial screen identified PAX8 as a driver of ovarian cancer[116].

Tumors rely on numerous processes for their survival, from uncontrolled growth, through evading cytotoxic signals, to changing their microenvironment to support their growth. Driver genes, therefore, are selected to support these various functions. Studies that use one discovery approach or focus solely on growth phenotypes are likely to miss many important drivers. The frequency, association and functional approaches are not mutually exclusive, and approaches that integrate input and insights of all these methods are needed to create a comprehensive view of cancer aberrations.

Drivers in melanoma

The use of sequencing and other high-throughput technologies drastically improved our understanding of melanoma genetics. Thus far, genomic and functional studies have implicated

several pathways in melanoma progression and survival, including MAPK[100], PI3K[117], KIT[118], MITF[119], TGFβ[120], WNT[121], CDKN2A[122], JAK/STAT[123] and several others[124].

Functional assays and clinical results, however, have focused most of the research on the MAPK pathway, which was shown to be a major driver of almost all melanomas. Here I give a brief introduction to the roles several major pathways play in melanoma progression, survival and drug resistance.

MAPK

The ERK-MAPK (Extracellular Regulated Kinase – Mitogen Activated Protein Kinase) pathway is constitutively active in ~70% of tumors. NRAS is mutated in ~20% of tumors, and BRAF harbors a point mutation (V600) in ~50% of tumors[125]. Studies show that MAPK activation is an early event in melanoma progression, as it is observed at similar frequencies in benign nevi and metastatic melanomas[56].

MAPK-activated tumors are highly dependent on MAPK activity. Inhibition of the pathway induces cell cycle arrest and cell death in many of the melanoma models, including cell lines, xenografts and patients[126,127].

MAPK influences a number of key cellular processes by regulating numerous pathways. MAPK controls the transition of cell cycle between G1 to S phases by negative regulation of p27[128], and positive regulation of both Myc[129] and Cyclin D1[130], among others. It inhibits cell death by inhibition of the pro-apoptotic protein BIM[131], and regulates cell growth and metabolism through activation of mTOR[132].

Due to the dependency of many cancer types, including melanoma, on MAPK activity, several drugs that inhibit the pathway were developed. Of notice is PLX4032, a highly potent inhibitor specific to mutated BRAF[133]. The drug has been recently approved for melanoma treatment and shows good clinical results[127]. However, not all tumors respond to the drug, and the molecular mechanisms underlying response heterogeneity are still not understood. Dual inhibition of MEK and BRAF in BRAF-mutated tumors demonstrates slightly better clinical results, but no survival benefit[28,134].

MITF

Micropthalmia-associated transcription factor (MITF) is a master regulator of melanocyte development, function and survival[119]. It promotes melanocyte differentiation by regulating the expression of genes required for melanin production and melanocyte differentiation from neural-crust pre-melanocyte cell to a differentiated and functional melanocyte.

In 2004 MITF was found to be amplified and overexpressed in 10-20% of melanomas[103]. Further studies have revealed a higher frequency of MITF overexpression in nevi compared with metastatic lesions[135]. It was also shown to predict survival, with low MITF correlating with worse outcome[136]. Additionally, in vitro studies found the MITF expression correlates with cell lines' response to MAPK inhibition[137].

Under physiological conditions MITF is regulated by α-MSH, a melanocyte-stimulating hormone. α-MSH binds to MC1R, and through the accumulation of cAMP they activate CREB. In turn, CREB stimulates the promoter of MITF-M, a lineage specific isoform of MITF[119]. Interestingly, germ-line mutation in MC1R are linked with higher incidence of melanoma[138,139]. Regulation of MITF in melanoma is more complicated. While MAPK and KIT can activate CREB and lead to MITF expression, they also negatively control production and accumulation of MITF protein[140-142].

Functional assays, unfortunately, do not help to clarify the role of MITF in melanoma. Knockdown of MITF in cell lines with high MITF expression leads to growth inhibition, while overexpression of MITF in other cell lines leads to terminal differentiation and lower tumoigeneicity[135].

Additional studies are required to understand the role of MITF in melanoma, and to determine whether it can be used as a biomarker or a target for therapy.

p16 and cell cycle proteins

CDKN2A, or p16, was first implicated in melanoma after identifying germ-line mutations in familial melanoma[122,143]. Mutations and loss of the gene locus were also described in sporadic

cases and cell lines[144,145]. p16 deletion and mutation are now reported in ~50% of melanomas (TCGA, *in press*).

p16 is a cyclin-dependent kinase inhibitor that blocks CyclinD to prevent G1 to S transition. It acts as a tumor suppressor and its loss disrupts cell cycle regulation. Other components of the cell cycle regulation are often deregulated in cancer. In melanoma these aberrations include loss of Rb1, Cyclin D1 overexpression or CDK4 amplification and overexpression[146,147].

Pharmaceutical inhibition of the cell cycle machinery is currently being tested. Although initial studies report low efficacy of CDK inhibitors, combinatorial treatments and second-generation inhibitors are still being tested[148].

PTEN and the PI3K pathway

The pro-growth anti-apoptosis PI3K-AKT pathway is activated in melanoma by PTEN loss or AKT3 amplification[149-151]. Among many of its targets are the growth regulator mTOR, CDK inhibitors, FOXO transcription factors, MDM2, the bcl-associated death promoter (BAD) and others[152].

PTEN loss was found by several independent studies to correlate with response to MAPK (MEK or BRAF) inhibition[42,153]. Supporting these studies are results that show that concurrent inhibition of the PI3K and MAPK pathways leads to stronger cytotoxic and cytostatic responses[154-156]. However, others have suggested that PTEN plays other roles in melanoma which are unrelated to proliferation and apoptosis[157]. Additionally, no correlation between PTEN status and response to treatment in patients has been reported, and the role of the PTEN-AKT pathway has yet to be fully determined.

Open questions

High throughput technologies greatly advanced our understanding on tumorigenesis as a whole, and melanoma progression and survival specifically. More recently, several new drugs, both immuno-modulators and kinase inhibitors, have been approved for melanoma treatment. However, initial clinical results demonstrate response heterogeneity. The molecular reasons underlying response heterogeneity are still unknown.

A better understanding on the molecular biology of melanoma is likely to assist with the pursuit of better clinical results. Although several large-scale high throughput studies aimed to identify oncogenes and tumor suppressor have been conducted, our knowledge is still lacking: The drivers within many of the frequent copy number aberrant regions have not been identified; The genetic and molecular determinants of phenotypic response to MAPK inhibition are still unknown; Combinatorial treatments show good in vitro results, but toxicity issues thwart clinical use of the combinations and better targeted therapy regimens are needed.

Chapter II - Context-Specific Interactions in the Genetics of Gene Expression

This chapter was published under Litvin et al. *PNAS* 2009[158]

Introduction

Understanding the effect of genetic sequence variation on phenotype is a major challenge that lies at the heart of genetics. Recent technological advances in genotyping have now made it possible to obtain a comprehensive view of genome-wide variation in a large number of individuals. However, studies that associate genetic polymorphisms with phenotypic properties (disease, height, etc.) involving tens of thousands of individuals[159] have, for the most part, only been able to detect loci that collectively account for 3% of the heritable phenotype. This suggests that the connection between genotype and phenotype is more complex than previously assumed and that more sophisticated approaches are needed to interpret the data.

The goal of this project is to elucidate the relationship between genotype and phenotype, and to map the landscape of genetic interactions. I use gene expression as a phenotype, and associates clusters of gene expression with genetic variance. Changes in cell state, such as metabolic state and growth rate, as well as activation of various pathways, are likely to affect the expression of many genes. In this approach, expression quantitative trait loci (eQTL), is based on the notion that gene expression reflects cellular state.

Quantitative trait mapping of gene expression abundances has proved a powerful model system for studying genetic traits in a number of organisms[15,160-162]. I use gene expression and genotype data on segregants generated in a cross between a laboratory strain (BY) and a wild strain (RM) of *Saccharomyces cerevisiae*[163,164].

I developed GOLPH (GenOmic Linkage to PHenotype), a novel statistical algorithm to identify multiple genetic factors influencing gene expression abundance. GOLPH's premise is that the modular organization of gene regulation can be used to enhance the statistical power of linkage to eQTLs, since clusters reduce the statistical burden of the test while also reducing the noise in the data.

GOLPH identifies an unprecedented number of linked regions for each gene. The results portray a complex picture in which multiple loci influence the expression of modules of co-expressed genes that define coherent biological processes. The data show that genetic polymorphism can give rise to distinct cellular states in which entire metabolic pathways and biological processes are activated to different extents between individuals. In this regard, genotypic differences are similar to environmental perturbations in their effect on the internal state of the cell.

Moreover, most interacting loci demonstrate allele-specific genetic interactions, in which the secondary locus exerts an influence only in a specific context of the primary locus. A possible explanation is that the primary locus switches the cell among states or predisposes it towards adopting a cellular state. The secondary locus only has an effect in one of these states. For example, I observe differences in the cellular state mediated by variation at the *IRA2* locus. Genetic variation in *IRA2*, an inhibitor of RAS/PKA signaling, predisposes strains with the *IRA2-RM* allele towards aerobic respiration[164]. Other loci containing genes with critical functions involved with mitochondria and respiration exhibit *IRA2-RM* specific influence on entire transcriptional programs.

To support the biological and statistical analysis, I developed and used GENATOMY, a custom-built analysis tool, to visualize and analyze the resulting genetic interactions between quantitative trait loci. GENATOMY is available for download on the lab website.

The data and results depict a complex relationship between genotype and phenotype resulting from the dynamic nature of genetic interaction networks that are responsive to both the environment and genetic variation.

Results

I developed GOLPH, a new statistical approach to find multi-locus linkage or association to gene expression traits. It is based on the detection of *iQTL* (interacting Quantitative Trait Loci) that involve two or three loci. Each iQTL consists of a primary locus and up to two secondary interacting loci, which significantly link to the trait in a context-specific manner - only when the primary locus has a specific allele. GOLPH constructs *iQTL modules* consisting of the iQTL decision tree and all the genes that link to that combination of interacting loci (Figure II-1). These iQTL modules are further partitioned into subsets of co-expressed genes, referred to as *expression patterns*.

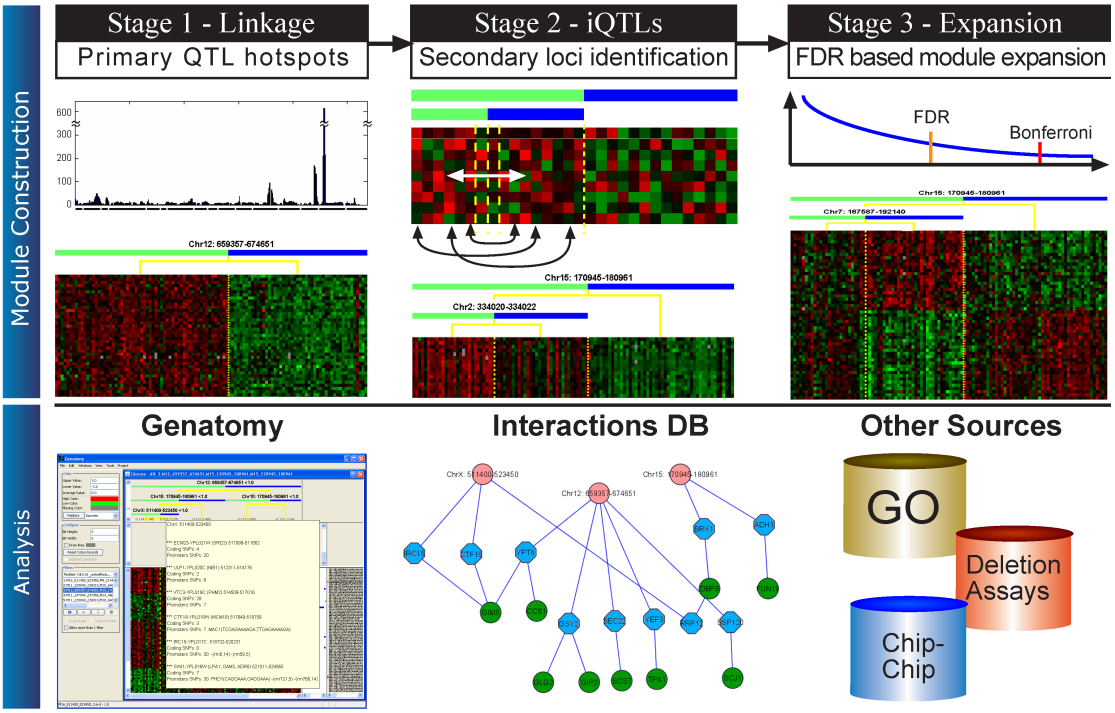


Figure III-1 - Overview of GOLPH
GOLPH takes as input gene expression and genotype data for a set of individuals. Top panel: shows the computation occurring at each stage and the resulting output (bottom panel). Stage one - genes are linked to a primary locus; stage two - iQTL are constructed by partitioning the samples based on the primary locus and linkage to the secondary locus; stage three -, FDR is used to expand significant linkages. Bottom panel: Once all iQTL modules have been constructed they are analyzed using GENATOMY, our interactive visualization and data analysis tool. GENATOMY uses additional resources such as sequence, Gene Ontology (GO) annotations, protein DNA interactions and genetic interactions to help interpret the data.

I applied GOLPH to genotype and gene expression data obtained from 108 segregants and their parents[15,163,164]. GOLPH works in three stages, each increasing the number of detected linkages. Similar to previous studies[17,165], GOLPH begins with a stepwise search. In the first stage, primary QTLs are detected for each trait, and in the second stage, secondary interacting loci are detected. In contrast to previous studies[165], a secondary QTL is identified independently for every allele at the primary locus. In the final phase I exploit the modular organization of gene regulation to link genes that are not significant alone, but which share a pattern with significantly linked genes (see Figure II-1 and *Methods*). I analyzed the resulting linkages using GENATOMY, a purpose-built visualization tool, to gain insight into the architecture of interacting loci.

The GOLPH algorithm significantly increases the number of linkages

Stage one of the analysis identified 44 hotspots including many previously reported regions (*AMN1*, *GPA1*, *HAP1*, *IRA2*, *MKT1, PHO84*)[15,162,164,166,167]. Using these hotspots, stage two

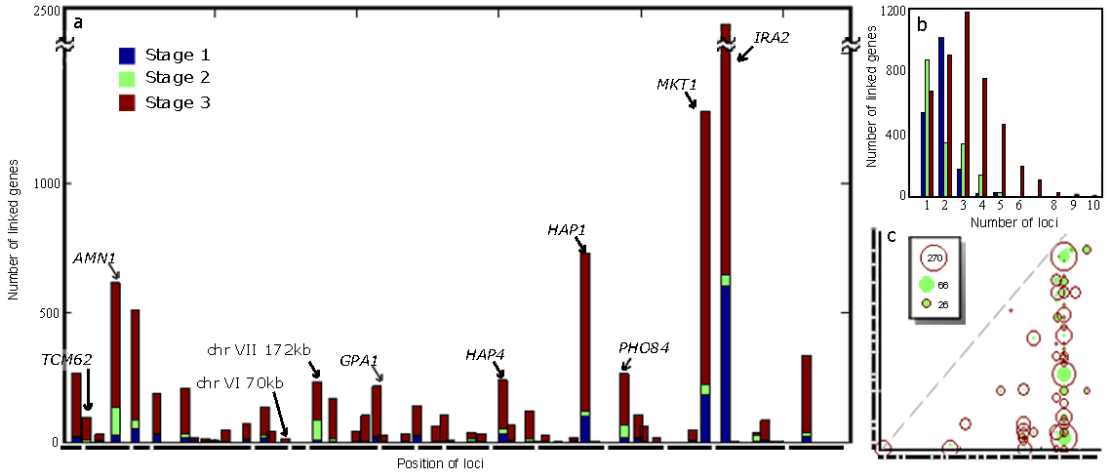


Figure III-2 – The genetic landscape of eQTL
The number of linked genes increases at each of the three stages, with the greatest expansion in module size occurring at stage 3. **(a)** Number of linkages at each locus color coded by stage 1 (blue), 2 (green) and 3 (red). The x axis represents the location of the locus, each of the bold lines below the axis represent yeast chromosomes I to XVI. The y axis represents the number of genes linked to that locus. **(b)** Histogram representing the number of loci linking to each gene at each of the three stages. The color code is the same as in 2(a). **(c)** Plot showing the size of each iQTL at stages 2 (green) and 3 (red). The size of the circle is proportional to the number of genes linked to the iQTL). Both axes relate to chromosomal location with the position of the chromosome marked in bold.

identifies secondary loci that interact with each of the primary loci. 81 pairs of iQTL that link to 5

or more genes were found, resulting in an increase in the number of multi-locus genes (Figure II-2a,b).

In stage three, GOLPH uses the modularity of gene expression to gain additional power. The premise is that gene regulatory networks are organized into modules of co-regulated genes[168,169]. For example, deletion studies have shown that when a regulator is deleted, the expression of hundreds of genes are influenced[170]. Therefore, weaker linkage of additional genes to the iQTL identified in stage two are more likely to be real. This leads to a dramatic increase in both the number of genes linked to each marker and the number of markers linked to each gene (Figure II-2). After stage three, more than 2500 genes linked to two or more loci and more than 800 genes linked to five or more loci, matching previous analysis estimating that the expression of more than half the genes is likely influenced by at least five different loci[163]. To ensure GOLPH does not report spurious linkage, randomization testing for each step is performed, and indeed no signal is detected for the randomized data (see Methods). I conclude that GOLPH detects an unprecedented number of loci for each gene expression trait and demonstrates that genetic interactions between loci are more common than previously estimated[166].

Interacting QTLs generate coordinated biological programs of gene expression

Although genes were added to each iQTL module based on their linkage alone, the resulting sets of genes form tightly co-expressed clusters. Figure II-3 shows the set of genes that are added to an iQTL module involving *IRA2* (Chromosome XV:170945-180961) and the chromosome VII locus (Chromosome VII:167587-192140) at each stage. The genes added in stage three have the same pattern of expression as the genes added during the more rigorous stage two. In addition, the genes added in stage three share Gene Ontology (GO) annotations and binding sites with those chosen in stages one and two, significantly improving the functional enrichment of the modules, and further supporting their linkage (see Methods). Examples of improved enrichment include ribosome biogenesis and assembly: 10^{-47} in stage two to 10^{-102} after stage 3, mitochondrion from 10^{-18} to 10^{-76}, iron ion transport 10^{-4} to 10^{-10} and aerobic respiration 10^{-3} to 10^{-12}. We conclude that iQTL do not influence a single gene but rather entire biological processes and pathways.

Figure II-3 shows two distinct patterns of co-expressed genes that are inverted, *i.e.*, down-regulation on one side of the heat map is accompanied by up-regulation of equivalent magnitude on the other side and vice-versa. This is a widespread phenomenon involving 122 modules and 3638 linked genes, resembling the response to environmental perturbation, in which entire processes are co-coordinately up- or down- regulated (see Methods). The existence of inverse expression patterns suggests that many iQTL not only regulate single pathways, but rather orchestrate entire cellular responses involving multiple biological processes.

My results provide a view of genetic variation as an internal cue that predisposes the organism towards, or away from a cellular state. The presence of a single allele can tip the balance between one state and another. The most striking example is provided by the *IRA2* locus which links to more than 2000 genes. Ira2 is a GTPase-activating protein that negatively regulates RAS. The RAS/PKA pathway plays a central role in coordinating processes such as growth and stress tolerance in response to nutrient availability. The *IRA2*-RM sequence differs from BY by 87 non-synonymous coding SNPs and 3 gaps.

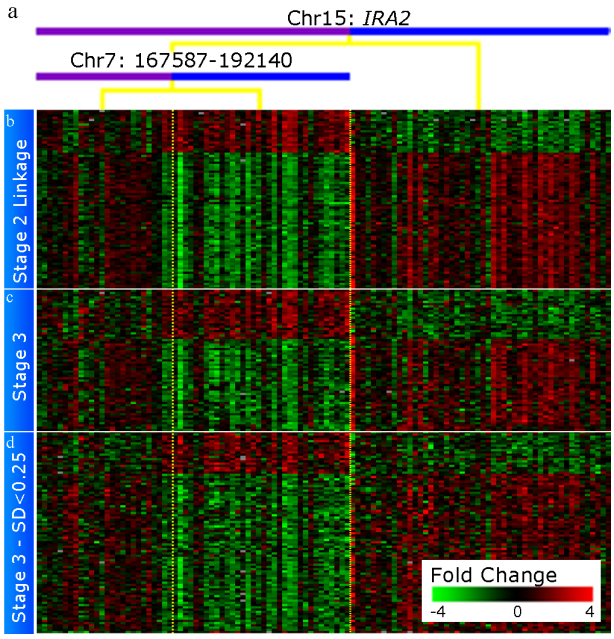


Figure III-3 - IRA2 module
A heat map showing the *IRA2*–chrVII iQTL module and the expression of the genes linked at stages 2 and 3. Each row represents a gene and each column represents a strain. The module is organized as a decision tree based on the strain's genotype and whether they inherited the BY (blue) or RM (purple) genotype for each of the interacting loci. **(a)** Top split based on the primary locus, chromosome XV:*IRA2.* The lower split is based on the secondary locus chromosome VII:167587-192140. **(b)** 80 genes linked in stage two. The columns represent strains and are arranged according to the tree, the vertical dotted yellow lines show the split point in the genotype. **(c)** 62 genes linked in stage 3. The variance in expression of these genes is >0.25 SD. These genes were considered in stages 1 and 2, but did not pass the higher threshold for significance. **(d)** An additional 88 genes are linked in stage 3. These genes are not considered in stage two because their variance in expression is < 0.25 SD and are hence noisier.

Segregants with the RM allele of *IRA2* correspondingly inhibit Ras/PKA signaling better than segregants with the BY allele[164]. Although all of the segregants were grown in glucose (and might

be expected to undergo fermentative growth), the presence of the RM allele correlates with the up-regulation of genes annotated for mitochondria (10^{-14}), aerobic respiration (10^{-9}), response to stress (10^{-8}) and the down-regulation of genes annotated for ribosome biogenesis and assembly (10^{-95}), rRNA processing (10^{-57}) and the nucleolus (10^{-56}), suggesting a transcriptional response consistent with respiration.

In contrast to *IRA2*, the phenotypic differences that link to the *HAP4* (Chromosome XI: 247944_247956) locus are likely to be driven by allelic differences in the promoter. Hap4 is part of a transcriptional activator complex that regulates the transcription of genes in response to heme/oxygen and/or growth on non-fermentable substrates [171] and the locus is linked to more than 200 genes. *HAP4* is a cis-eQTL, *i.e.*, a gene that links to its own locus, and the RM strain has 14 promoter SNPs. Moreover, the presence of the *HAP4-RM* allele correspondingly correlates with the up-regulation of *HAP4* along with genes it activates: Hap4 bound genes (10^{-19}), those annotated for mitochondria (10^{-90}), and aerobic respiration (10^{-13}).

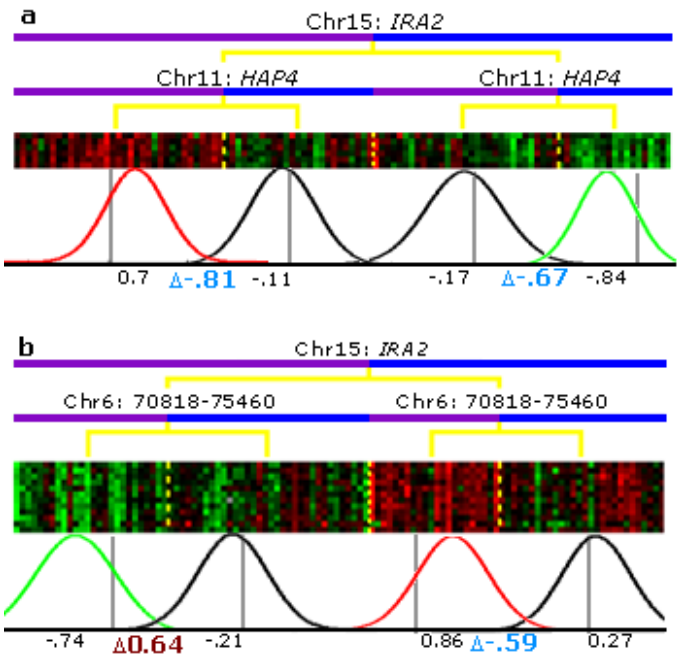


Figure III-4 - Linear and Non-linear iQTLs
(a) An example of an iQTL module with an additive interaction between the *IRA2* primary locus and the *HAP4* secondary locus. The colored number (red for positive and blue for negative) represents the difference in means of gene expression between each side of the split. We observe that the delta between *HAP4*-BY and *HAP4*-RM segregants is similar for both *IRA2-BY* and *IRA2-RM* ; hence the interaction between these loci is additive. **(b)** An example of an iQTL module with an opposing interaction between the *IRA2* primary locus and chromosome VI:70818-75460. The delta between chrVI-BY and chrVI-RM is -0.59 in the context of *IRA2-BY* and reversed (0.64) in the context of *IRA2-RM.*

The landscape of genetic interactions

GOLPH detected 83 pairs of interacting loci in 205 modules with 542 expression patterns. I used the multi-locus phenotypes to characterize the genetic interactions between QTLs. Most methods for multi-locus traits assume an additive model, y~aX+bY. For example, the iQTL module involving the *IRA2* and *HAP4* loci influence different aspects of mitochondrial function. The *IRA2-RM* allele represses the PKA pathway, predisposing the strain towards respiratory growth. Hap4, an activator of aerobic respiration is upregulated in segregants with the *HAP4-RM* locus. Therefore, the presence of *IRA2-RM* and *HAP4-RM* each push the cell towards respiration through independent mechanisms and their joint influence is an additive combination of their individual influences (Figure II-4a).

One of the most striking aspects of the data is the dominance of allele-specific interactions, *i.e.*, situations in which the secondary locus exerts an influence on the phenotype only when the primary locus has a particular allele (and has little or no influence when the primary locus has another allele). GOLPH is able to detect allele-specific interactions because each primary allele is tested for linkage independently and the secondary locus need not link to both. One such example was presented in Figure

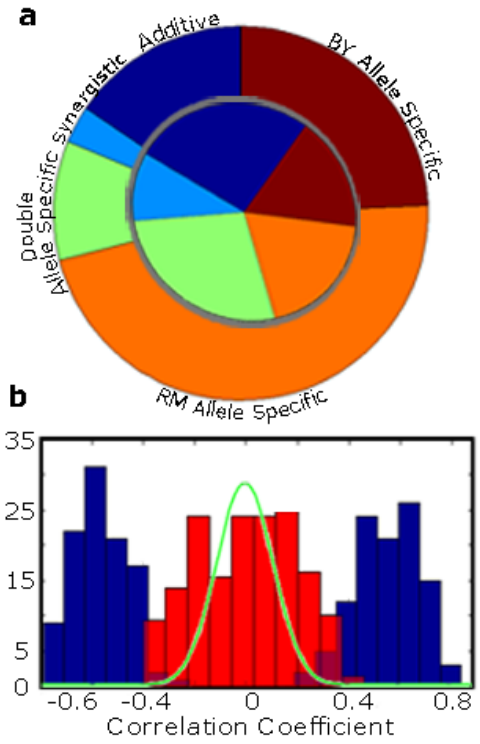


Figure III-5 - The landscape of iQTLs
(a) A pie chart representing the types of interactions between loci in our analysis. The outer circle represents genes and the inner circle represents modules. RM allele-specific interactions are orange, BY allele-specific interactions are brown. Blue represents situations in which the secondary allele links to both sides, additive interactions are dark blue and synergistic interactions are light blue. Green represents modules with two different allele-specific interactions, one for each side. The dominance of allele-specific interactions is evident. **(b)** Histogram of correlation coefficients in allele-specific modules. The data show that the effect of the secondary locus on the non-interacting allele is negligible. The x axis is the correlation coefficient between the secondary locus and the mean expression level for genes in the module. The y axis shows the number of modules. The blue bars represent data from the interacting primary allele and the red bars represent the other non-interacting allele. The green line shows that the distribution for randomly chosen pairs of loci is similar to the histogram in red demonstrating that the interactions are indeed with only one allele and not the other.

II-3 - the chromosome VII locus interacts with *IRA2-RM* only and has no influence on *IRA2-BY*. While 196 genes link to the chromosome VII locus in with the presence of *IRA2-RM*, none of these linkage signals were significant in stage one. To confirm that these interacting loci are indeed allele-specific and do not reflect borderline effects, I compared the regression coefficient of the linked versus non-linked alleles and found that coefficients for the non-linked alleles resemble a random distribution (Figure II-5b). GOLPH detected a remarkable number of allele-specific genetic interactions. These involve 78 interacting loci, organized into 94 iQTL modules that contain 1856 unique genes and 2891 allele-specific interactions, 81% of the total interactions identified (Figure II-5a). I conclude that allele-specific genetic interactions are prevalent in our data.

The same secondary locus was found to influence both primary alleles in 50 iQTL modules containing a total of 562 genes. I tested each of these for epistasis (see Methods), and in cases where the secondary locus links to both primary alleles, the majority of interactions (423/562) do not show a significant interaction term. Since most other methods do not detect allele-specific interactions, this could explain why genetic interactions are typically assumed to be additive. While there are only a few epistatic modules, these can exhibit dramatic effects; a number of iQTL modules had secondary locus effects in opposing directions between the two primary alleles. For instance in the iQTL involving the *IRA2* and chromosome VI:70818_75460, the effect of the chromosome VI locus depends on the *IRA2* allele. The RM allele of chromosome VI:70818_75460 up-regulates the genes in the module in the presence of *IRA2-BY* and down-regulates the genes in the presence of *IRA2-RM* (Figure II-4b).

The prevalence of allele-specific genetic interactions

To understand how allele-specificity might arise, I analyzed an iQTL module linked to *IRA2-RM* and a locus on chromosome II: 334020_334022 (Figure II-6). The causal gene on chromosome II is likely to be *TCM62,* which encodes a protein that supports biogenesis of the mitochondrial succinate dehdrogenase complex by acting as a molecular chaperone [172]. Strains deleted for *TCM62* grow slowly on rich glycerol medium and are respiration deficient. *TCM62-RM* has 3 coding SNPs and 57 promoter SNPs compared with the BY sequence, including SNPs in

two Pho2 binding sites. One of these SNPs is predicted to increase the binding affinity of Pho2 [173]; indeed, *TCM62* is a strong cis-eQTL and is up-regulated in segregants bearing the *TCM62-RM* allele.

When segregants have both *IRA2-RM* and *TCM62-RM,* mitochondrial genes (10^{-7}) and Skn7 targets (10^{-5}) are up-regulated, while ribosome biogenesis and assembly (10^{-63}), nucleolar (10^{-37}) and rRNA processing genes (10^{-32}) are down-regulated. Figure II-6 shows the expression pattern of genes in the module and that the *TCM62* locus has a strong influence in segregants with the *IRA2-RM* allele and negligible influence in segregants with the *IRA2-BY* allele. The PKA pathway is inhibited in segregants bearing the *IRA2-RM* allele and the balance is tipped towards the expression of genes associated with respiratory growth and the up-regulated *TCM62-RM* allele likely further tips the cell towards respiratory growth.

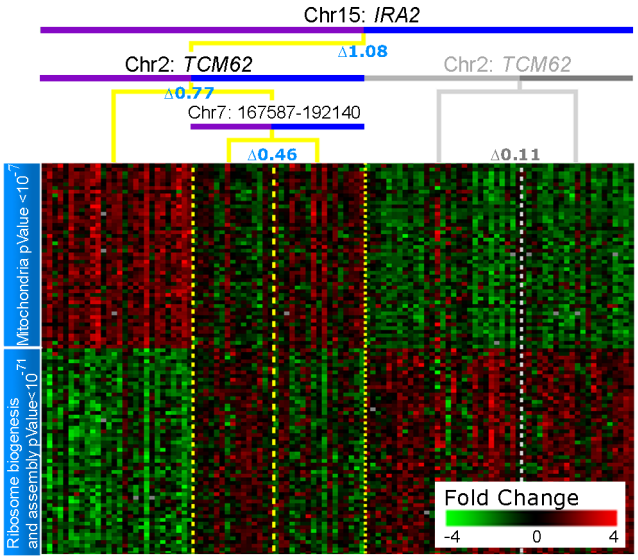


Figure III-6 - IRA2 module
The *IRA2–TCM62* iQTL module is graphically represented as described in Figure II-3. For compactness, representative genes were chosen for each pattern. The full list of genes for each pattern is provided in Supporting (delta 0.11 is low even by random scales). We manually added an additional partition using the chromosome VII locus from Figure 3 to *TCM62-BY* to demonstrate that the chromosome VII locus represents an alternative pathway to that affected by *TCM62-RM*.

Previous work reports that linkage to a particular locus is often dependent on environment, the locus exerting an influence in one environment, but not in another[164]. Both external environmental signals and genetically driven internal cues can drive cells to switch between states, as reflected by different metabolic fluxes and stresses acting on the cell. I postulate that the allele-specific linkages we detect are largely due to such events. These switches can sometimes be subtle, such as a release of inhibition or a shift in bottlenecks, making certain genes more critical in some conditions than

others. Thus genetic variation leads to internal change, altering interactions between genes through shutdown or activation of pathways, release of inhibition, or shifting of bottlenecks.

The *IRA2-RM* allele plays a dominant role among the allele-specific iQTL modules accounting for 41% of the genes influenced by allele-specific interactions with RM; however, other loci also exhibit this phenomenon. Each of *HAP1-RM* and *MKT1-RM* has allele-specific effects on the expression of more than 200 genes. Although there are fewer allele-specific interactions with BY, 886 genes are affected by allele-specific interactions on the BY side. The loci that dominate BY allele-specific interactions include *HAP1-BY* and a locus on chromosome I:41483_42639, likely to be due to polymorphism in *OAF1*, an oleate-activated transcription factor involved in the beta-oxidation of fatty acids and peroxisome organization and biogenesis. Together there are more than 10 different hotspots that exert allele-specific influences over a large number of genes.

Materials and Methods

Data

The strains, genotypes and gene expression measurements were those of [164]. I merged adjacent, highly-correlated markers, to obtain a total of 526 markers [174]. Expression data were normalized with mean of zero and variance one. For stages one and two of GOLPH, only 1733 genes that showed significant variation (stdev>0.25) in their expression level were used. GO categories from http://www.yeastgenome.org/ with more than 5 genes were used for the evaluation of biological function. Putative transcription factor binding sites were obtained from the Fraenkel lab web site http://fraenkel.mit.edu/yeast_map_2006/.

GOLPH algorithm

GOLPH aims to find multi-locus linkage or association to gene expression traits and is designed to find iQTL (interacting Quantitative Trail Loci) that involve two or three loci. Each iQTL consists of a primary locus that links to the trait and up to two secondary interacting loci. A secondary allele is included if it significantly links to the trait, conditioned on a specific allele for the primary locus. An iQTL can have a different secondary locus for each allele of the primary locus. The relationship between loci is represented as a decision tree with the primary locus set as the root node of the tree. GOLPH constructs iQTL modules consisting of the iQTL decision tree and all the genes that link to that combination of interacting loci. These iQTL modules are further portioned in subsets of co-expressed genes, or gene expression patterns.

GOLPH performs three non-iterative stages. First, it detects primary QTL and creates modules of all the genes that link to each locus. The second stage detects secondary loci and represents these as a decision tree, or regulatory program, and the third stage reassigns genes to iQTL modules based on the regulatory programs defined in stage 2. Stages 1 and 2 include only genes with standard deviation > 0.25 and stage 3 expands to all genes in the data.

At the heart of GOLPH is the modularity assumption. Stated simply, due to the modular organization of the regulatory network, true QTLs are likely to influence the expression of many genes. It is widely accepted that the cell's regulatory network has evolved to be modular,

reflecting common biological processes and pathways sharing regulatory mechanisms[168]. Indeed, it has been shown that the deletion of a single regulator alters the gene expression of hundreds of genes: Transcription factors[170,175], signaling molecules[176] and chromatin modifiers[177]. Moreover, it has been shown that most transcription factors directly bind hundreds of genes[178]. Therefore, if gene deletion alters the expression of hundreds of genes, it is reasonable to expect that other genetic changes which alter a regulator's function or abundance should also have significant influence on many transcripts.

Stage 1

Stage 1 is similar to classic linkage methods that are based on non-parametric permutation testing[15,179]. Each gene is tested against each locus and significant associations are reported. After associating genotypes to genes, I cluster nearby small linkage peaks and merge them with the largest dominant linkage peak along the chromosome (Figure II-7).

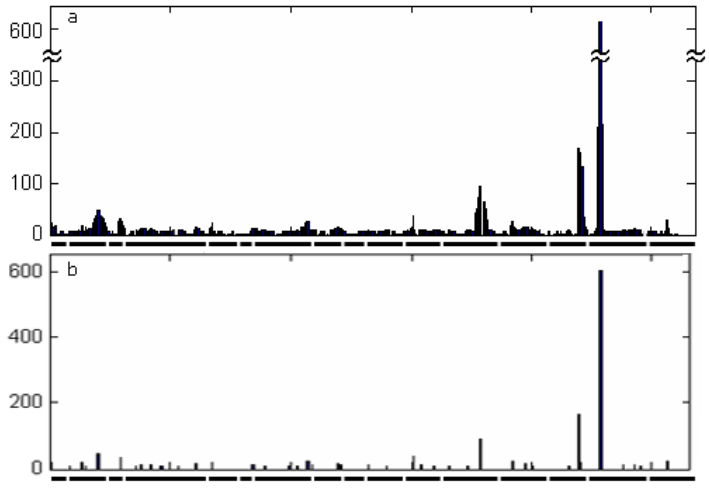


Figure III-7 - Merging close loci
(a) The top panel shows the linkage signal before unifying signals from nearby markers. The signal creates broad hills, rather than sharp peaks. The x axis represents the location of the locus, each of the bold lines below the axis represents a yeast chromosome (I to XVI). The y axis represents the number of genes linked to each locus. **(b)** The bottom panel shows a much cleaner linkage map with sharper peaks after unification step.

To detect significant gene-locus pairs, GOLPH uses Welch's t-test[180] splitting the segregants into two groups based on their genotype (BY or RM). Welch's t-test is an adaptation of the Student's t-test intended for use with two samples that may have unequal variances. This is similar to regression commonly used for association testing because the segregant data is binary; a gene can only be inherited from a BY or RM strain.

GOLPH evaluates the significance of each gene-locus pair via two tests: the parametric p-value of the Welch t-test and non-parametric permutation testing. For each gene 1000 random

permutations of segregant label were used, randomly assigning each gene expression profile to the genotype of another segregant. This way both gene expression and genotype hold their natural structure, but are independent of one another. All 526 markers for all 1000 random permutations were scored, generating a null distribution of Welch's t-test p-value for each gene. A gene is linked to a locus if its Welch p-value is better than 0.05 and permutation based p-value is better than 10^{-5} (p-value of 0.001 after Bonferroni correction for 526 markers).

Genes linked to one marker are also likely to have linkage signals in neighboring markers due to similarities between proximal markers (Figure II-7a). To pinpoint the most likely marker, I merge small peaks with proximal larger peaks. Based on modularity, the assumption is that if a nearby marker linked to a large number of genes, it is more likely to involve a causal factor than one that links to only a few genes. For each marker I construct a gene set including all the genes that link to that marker. I remove genes from small gene sets and move them to larger nearby loci based on the ratio between the size of the linked sets and the delta p-value of the linkage score. I run the algorithm several times, each time handling a different order of linkage peaks to get a stable linkage map. After removing modules that have fewer than five genes, 44 loci remain, each linked to at least five genes (Figure II-7b and Table II-1). Only two adjacent loci remain in the final map, each linked to a different set of genes.

Table III-1 - Module Growth

Locus	# genes linked after stage 1)	# of genes linked after all stage, out of 1733 genes	# genes linked after all stage out of all genes
M15_170945_180961	604	990	2128
M14_449639_449639	166	402	1088
M12_659357_674651	92	302	697
M2_533262_555787	47	128	292
M11_247944_247956	36	172	273
M3_100213_105042	31	101	173
M16_511400_523450	26	216	390
M13_46070_69122	24	82	181
M8_111679_111690	23	67	209
M5_420595_430931	22	76	123
M14_486861_486861	20	368	1018
M1_41483_42639	20	130	210
M2_310928_310928	19	154	282
M13_261719_286122	16	40	110
M9_139462_141014	15	40	63
M5_109310_117705	14	63	114

M12_677957_697260	12	190	407
M4_246738_262796	12	17	31
M15_838599_850119	11	20	39
M9_214482_254745	11	130	221
M7_98231_117900	10	9	17
M4_555043_555043	10	13	20
M15_515917_546197	9	9	26
M11_388373_397458	9	10	11
M10_163850_185319	9	9	14
M8_185012_188851	9	17	32
M13_404546_404546	8	40	66
M12_423789_423789	8	10	16
M11_566015_566015	8	50	113
M10_472146_503466	8	11	27
M7_402833_415585	8	22	45
M4_1418647_1418647	8	21	67
M4_114155_122293	8	76	147
M2_562409_570229	8	132	321
M14_220948_220948	7	7	22
M12_92674_95639	7	16	30
M7_1058947_1063841	7	43	108
M4_1240155_1240245	7	6	10
M2_352257_368060	7	82	171
M15_968429_970605	6	5	7
M7_916471_919654	6	39	93
M4_935079_935079	6	8	17
M2_328489_334016	6	146	311
M2_87845_87845	6	2	2

Lists the number of linkages to the 44 primary loci selected in stage 1, throughout the stages of the algorithm. The first column lists all loci of stage 1 and the number of genes linked. After merging of close loci and removal of modules with less than 5 genes, 44 modules remain. The second and third columns list the number of linked genes in stage 3.

Stage 2

In the second stage, GOLPH constructs iQTL modules by finding secondary loci that interact with the primary loci identified in stage one. This approach is similar to that of Storey et al.[165], and they have demonstrated that this two-step approach outperforms the exhaustive 2D scan, which tests all pairs of loci. More power is gained by evaluating significantly fewer hypotheses.

For each of the 44 modules identified in stage 1, segregants are partitioned into two sets based on whether they are BY or RM at the primary locus. For each gene in the module, (i.e., each gene linked to the primary locus), secondary linked loci are searched using the Welch permutation test as described for stage 1, based on a Welch's t-test p-value of 0.05 and a permutation p-value threshold of 10^{-4}. This process is carried out independently for segregants

that have the BY or RM allele at the primary locus. Each detected secondary linkage creates an iQTL represented as a decision tree. The resulting tree can have secondary splits on the BY (right) side, the RM (left) side or both.

As in stage 1, close loci link to overlapping sets of genes. To pinpoint iQTL peaks and remove redundant iQTL modules based on neighboring loci, I use a different approach from the one used in stage 1. For each chromosome, I create a graph where each chromosomal marker is represented as a node. An edge exists between two nodes if there is a gene linked to both markers. I collapse each fully connected clique (a subgraph of nodes fully connected to each other by edges) into the single marker with the most linkages in that clique, resulting in fewer loci with more genes linked per locus to each other. I use the resulting markers to construct the iQTL modules, their regulation trees and the set of linked genes. After removing modules that have fewer than five genes, we obtain 91 iQTL modules.

Stage 3

In this stage iQTL modules are expended based on the modularity assumption. Using the regulatory programs found with a stringent statistical threshold in stage 2, genes are reassigned to modules using false discovery rate (FDR[181]) on the permutation based p-values. In loose formalism, one can view stage 2 of the algorithm as constructing a "prior" on iQTL, assigning a higher "prior probability" for iQTL that have strong linkage to at least 5 expression traits. Stage 3 recalculates linkage allowing this "prior" to weigh in (this is not technically a prior, as the data itself is used to construct it), leading to the inclusion of many additional genes in each iQTL module.

Each regulation tree is examined, and all 4338 genes are evaluated for each tree, involving two independent tests, depending on the structure of the tree. A gene has to pass both tests to be included in a module.

- If a tree has only one secondary-split (Figures I-2,5), both the major locus and the secondary locus are tested.
- If a tree has two secondary splits, both secondary splits are tested, and the primary locus is used only to partition the segregants.

Using permutations, a p-value distribution was independently generated for each of the two tests above. An FDR threshold is determined from the observed p-value and a final gene set that passes an FDR threshold of at least 0.01 on both tests was chosen. Hence the threshold is adaptive to the number of genes and the strength of linkage signal for each locus, so a large number of weak signals that point to the same locus increase the significance. As a result, a larger number of genes are added to the modules (Table I-1). In the case of two sided modules and in contrast to stage 2, a gene does not need to link to the primary locus to be tested for an iQTL module. It is possible for a gene to join a module without significant linkage to the primary locus as clear separation of expression patterns occurs only in the secondary split.

Modules sharing the same primary locus are often found to have overlapping gene sets, but one has an allele-specific interaction on the BY side, and the other has one on the RM side. In many cases, a regulatory program with these two sub-splits was not created in stage 2 due to the rigorous cutoff and the use of fewer genes. A unified regulatory program is now created, and the overlapping genes are removed from their single-sided modules. The new expanded regulatory programs are now tested again for linkages using the method in stage 3.

Expression Patterns

Linkage for modules is performed independently for each gene ignoring coexpression. Although no restriction on correlation of expression is forced by GOLPH, expression between genes in each module does correlate, and in most cases the majority of the genes in the module are captured by a small number of expression patterns (see Figures I-3,5). Typically, when more than one expression pattern exists in an iQTL module, the patterns always reversed, or mirrored. While one group of genes is up-regulated in one pattern, the other group is down-regulated, and vice-versa.

Statistical Validation

To test the credibility of our algorithm I performed a number of randomization tests to ensure that the reported linkages are rarely the result of spurious linkages.

Randomized Data

The dataset was randomized by permuting the labeling of the segregants in the gene expression data. Both gene expression and genotype datasets remain intact, preserving the modularity and co-expression in the gene expression matrix as well as the genomic organization of the genotype data. The difference is that the gene expression of a strain does not match the genotype and therefore any linkage is spurious. This is equivalent to the randomization used to evaluate Welch's t-test, except that here it is treat it as the “input data” for GOLPH.

I ran stage 1 on several permutated datasets and found only a small number (2-3) of small linkage peaks. This is compared to the original dataset in which many linkages including sets as large as 604 genes were detected. Stage 2 did not find any single iQTL with 5 or more genes. Therefore I conclude that GOLPH’s iQTLs are robust and represent true biological signal.

Randomized iQTL

Another concern is that by lowering the threshold, stage 3, which reassigns genes to pre-defined iQTL, might add many spurious linkages. To test my assumption that only true interacting loci will result in a large module (many genes) in stage 3, I randomized the pairs of iQTL given as input to stage 3 (Figure II-8). Randomizations were performed in two ways:

- Double random: 1000 random pairs of loci out of the available 526, where both the primary and secondary loci were selected at random, as well as a random determination of the BY or RM allele for the secondary split.
- Secondary random: 1000 random pairs of loci, where the primary locus is randomly chosen out of the 44 true primary loci and the secondary split is chosen at random from all 526 markers, as well as a random determination of the BY or RM allele for the secondary split.

Figure II-8 shows us that large iQTL modules do not form with random loci pairs, but rather with true interacting pairs. It is also evident why a requirement for at least five genes in the iQTL was used before applying the expansion: many random iQTL have a few genes linked to them, but rarely do random pairs have 5 or more genes. Among the 1000 randomized secondary loci, only 218 have more than one gene linked to them. 20 out of these 218 programs were identified by GOLPH from the original data, explaining linkages for 185 genes. Taking this into account, we found only 144 genes linked to the rest 198 modules by chance.

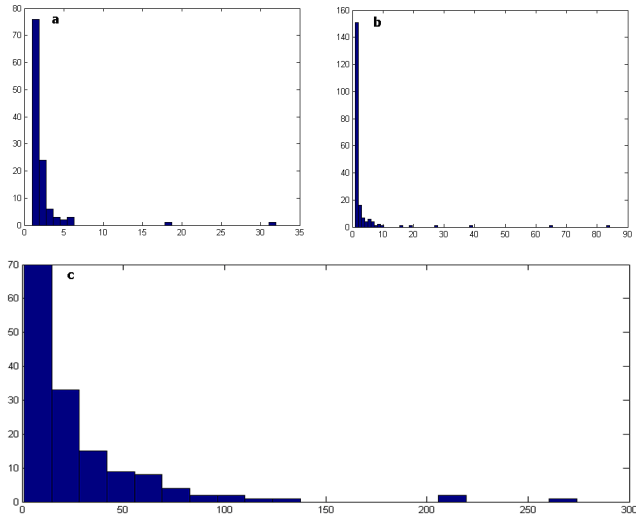


Figure III-8 - Results of Randomized iQTLs
Randomized allele pairs. The resulting module sizes (number of genes) in the x axis and number of modules for each size (out of 1000 random iQTLs) in the y axis. (a) represents the 1000 double random iQTL and (b) represents 1000 secondary random iQTL. (c) shows the results of the real data.

Taking these results together, I conclude that the iQTL chosen in stage 2 are statistically and biologically meaningful.

Validation using other data sources

The biological relevance of each iQTL module was evaluated using gene set enrichment analysis from a number of other genomic data sets and databases. GO categories with more than 5 genes were used to evaluate the functional coherence of modules. I compiled a number of resources to generate gene sets associating regulators and their targets. These include ChIP binding data for transcription factors[182] and chromatin modifying factors[183,184], gene expression data from deletion mutants[17,185] and putative transcription factor binding sites (TFBS) from the Fraenkel lab web site http://fraenkel.mit.edu/yeast_map_2006/ (The TFBS map with sequences conserved in at least two other yeast species and matched with a ChIP binding event was used).

Hypergeometric enrichments were calculated for all modules against all gene sets and carried out an FDR correction considering values of Pcorrected < 0.005 to be significant.

Definition of Interaction Types

Let X be the primary locus and Y, Z be secondary loci. GOLPH's model can be written as expression y=baseline+aX+αbY+(1-α)cZ, α=1 for X=BY and α=0 for X=RM.

Using the resulting model I define 4 interaction types. While the categorization is defined algorithmically and is based on the module structure identified by GOLPH, the resulting categories exhibit distinct characteristics.

Allele specific: Modules with only one secondary locus, linking to only one allele of the primary locus. In allele specific modules, the secondary locus is not significant for the other allele of the primary locus. This splits into two additional categories based on the primary allele, RM (Figure II-5a orange) and BY (Figure II-5a brown). A gene-locus pair (G,L) is considered allele specific if G resides in an allele specific module, with L being its secondary locus. I note that if G is included in a module that is two-sided for L, it is excluded from any allele specific module involving L as a secondary locus. This definition does not formally test that an independent contribution of L to G's expression is 0; nevertheless, Figure II-5b demonstrates that c is small and that c<<b for the case of BY allele specific and likewise b is small and b << c for RM allele specific.

Double Allele specific: Modules with two different secondary loci regulating each allele of the primary locus. Neither secondary locus links to the other side (Figure II-5a green).

Additive: Modules with the same secondary locus significant for both alleles of the primary locus (Figure II-5a dark blue). Additionally, F-test was used as described in[163] to test for synergistic interaction - whether the expression is better described as y~AX+BY or y~AX+BY+CXY. For additive modules, C is insignificant or b≈c in our model's formalism. An example of such a module is given in Figure II-4a.

Synergistic: Modules with the same secondary locus significant for both alleles of the primary locus (Figure II-5a light blue) that are best represented as y~AX+BY+CXY, i.e., C is a significant term in the F-test. In the model's formalism this means that b and c are significantly different and

can at times be even opposing in their effect. An example of such an opposing influence is given in Figure II-4b.

Fifty non-allele specific modules were identified by GOLPH, with a total of 562 genes creating 151 gene expression patterns. 31 patterns in 18 modules were found to have a significant interaction term, passing a p-value threshold of 0.05 in the F-Test (Table I-2).

Five out of the 31 synergistic programs show reverse effect of the second regulator. In these cases, the second regulator has one effect on the segregants with a RM allele of the major regulator, either over or under expression, while showing the reverse effect on the BY segregants (Figure II-4b, Table II-2).

Table III-2 – eQTL reverse interactions

Primary Locus	Secondary Locus	# Genes
M15_170945_180961	M7_167587_192140	1
M15_170945_180961	M4_161196_165032	1
M15_170945_180961	M10_89097_96376	1
M15_170945_180961	M9_190794_205191	1
M15_170945_180961	M4_85846_106892	2
M15_170945_180961	M4_85846_106892	1
M15_170945_180961	M4_85846_106892	1
M15_170945_180961	M12_92674_95639	1
M15_170945_180961	M10_472146_503466	1
M15_170945_180961	M10_472146_503466	6
M15_170945_180961	M9_21454_21455	1
M15_170945_180961	M9_21454_21455	26
M15_170945_180961	M9_21454_21455	1
M15_170945_180961	M9_21454_21455	3
M15_170945_180961	M16_84943_104423	3
M15_170945_180961	M12_450041_508029	1
M15_170945_180961	M12_450041_508029	1
M15_170945_180961	M6_70818_75460	1
M15_170945_180961	M6_70818_75460	3
M15_170945_180961	M6_70818_75460	12
M15_170945_180961	M3_100213_105042	3
M14_449639_449639	M12_642137_644136	1
M14_449639_449639	M1_51324_52943	1
M14_449639_449639	M1_51324_52943	2
M14_449639_449639	M1_51324_52943	4
M12_659357_674651	M15_802724_819015	1
M12_659357_674651	M5_492125_504717	3
M12_659357_674651	M5_492125_504717	2
M12_659357_674651	M1_41483_42639	1
M2_533262_555787	M13_33501_33681	2
M2_533262_555787	M13_33501_33681	1

31 modules that were identified as synergistic using F-test. Modules with reverse effect are highlighted

Comparison to previous studies

Previous studies by Storey et al.[165] developed an algorithm to detect interacting QTL using a stepwise algorithm. While both methods use a stepwise approach to identify secondary linkage, there are three key differences between GOLPH and these previous studies: modularity, allele-specificity and "one gene-one program" assumption.

Modularity

At the heart of GOLPH is our “modularity assumption" and I apply this assumption in a number of points during the algorithm. Stage 1 uses the assumption of modularity to unite adjacent markers which show linkage to overlapping sets of genes. Stage 2 only considers significant modules with 5 or more genes, and stage 3 uses modularity to reassign a higher “prior probability” to iQTL detected in stage 2. GOLPH uses the modularity assumption to lower the threshold of stage 3 by calculating FDR on sets of genes. The largest gain in GOLPH linkages (red bars in Figure II-2a) is due to this application of modularity in stage 3.

Allele-Specificity

GOLPH tests the secondary locus for each primary allele independently and can detect two different secondary alleles. The mathematical model assumed by Storey et al. can be written as: $y \sim aX+bY+cXY$, whereas GOLPH’s model can be written as $y \sim aX+\alpha bY+(1-\alpha)cZ$. The second and third genotypes (Y and Z) only affect the expression level when the primary locus (X) has a specific allele, hence the term "allele-specific iQTL."

"One Gene-One Program"

The above-mentioned studies looked for the single best primary linkage for each gene in the first step, and then used this best linkage to find a single best secondary linkage. My approach is not based on the assumption that only the best linkage found is true, but rather all linkages that pass the defined significance thresholds are considered to be true. This approach allows me to find more than two QTL for one gene. Figure II-2b demonstrates the number of linkages achieved for each gene and these match the theoretical studies by Brem and Kruglyak[163] on the expected number of influences for each gene.

As a consequence of these differences the linkages resulting from each of the two methods are different and complementary to each other. In total, GOLPH identifies many more linkage pairs (and single linkages) than the method of Storey et al. Many allele specific and opposing interactions are detected only by GOLPH; moreover in stage 3 many genes are added that do not pass significance by Storey’s FDR. The largest overlap in the detected linkages between the two

methods is in the interactions we characterized as additive. Nevertheless, Storey's method detects many additional additive linkage pairs that are missed by GOLPH. GOLPH removes from consideration all iQTLs involving fewer than 5 genes linking to that pair (motivated by our tests on randomized data) and for an individual gene versus a locus pair, the FDR implemented by Storey et al. is more powerful and sensitive.

Discussion

The emergence of new technological advances in high throughput genotyping and sequencing has enabled large scale characterization of genetic variation at high resolution. However, novel computational approaches are needed to detect causal sequence variants and model how genotype influences phenotype. A first step is to characterize the landscape of genetic interactions between naturally occurring variants and to elucidate how multiple loci combine to affect phenotype.

Applying GOLPH to yeast detected between two to ten linkages for each of 2745 genes, providing a first expansive view on the architecture of multi-locus traits and the genetic interactions between them. A remarkable finding is a large-scale occurrence of allele-specific interactions, indicating that the landscape of multi-locus traits is predominantly non-additive. A likely mechanism for allele-specific interactions stems from the observation that genetic variation can mimic the response to environmental change. Thus different biological states occur not only in response to the external environment but also as a result of intrinsic genetic variation.

Genetic variation in both coding and regulatory regions of transcription factors can lead to responses that alter cellular state (e.g. *HAP1, HAP4*). More intriguing, such large scale transcriptional responses are not only caused by variation in classical transcriptional regulators, but also due to polymorphism in metabolic enzymes, regulators of translation and molecular chaperones (e.g. *LEU2, MKT1, TCM62*). These demonstrate that genetic variation in a single gene may trigger a cascade of events, leading to an alternative cellular state, by predisposing the cell towards shutdown or activation of pathways. In this way the molecular network can be considered an intricate web of interacting factors in which dynamic entities may rewire their connectivity in response to perturbations in the environment and as a result of intrinsic genetic variation.

The prevalence of complex, non-additive gene-gene interactions is likely to play a large role in human and disease- related genetics and offers clues as to why recent association studies involving over tens of thousands of individuals have only accounted for a very small fraction of the heritable variation observed [159]. I believe that state changes driven by intrinsic genetic variation

and the resulting allele specific interactions are likely common in human and disease associated genetics. In multi-cellular organisms, genetic variation can lead not only to an altered cellular state, but can propagate to changes at the level of the entire organism. Detecting such allele-specific association in human is significantly more challenging as the genome is two orders of magnitude larger than yeast and the population structure is more complex.

This work also demonstrates the value of using gene expression as a proxy between genotype and phenotype. Contrary to previous association studies, I used clusters of genes and not single genes as the phenotypes. Aggregating genes in clusters proved to enhance the power of the study by two fold – First, clusters of genes cancel the noise in the data, allowing for more accurate association; Second, post-analysis on the cluster genes allow linking the loci to molecular functions, greatly enhancing the biological interpretation and value of linkage analysis. Others have shown that gene expression can also identify the casual gene within the loci[186], using gene expression both to predict a phenotype and as the phenotype itself.

Chapter III - An Integrated Approach to Uncover Drivers of Cancer

This work was done in collaboration with Dr. Uri-David Akavia and was published under Akavia, Litvin et al.[12]

Introduction

Large-scale initiatives to map chromosomal aberrations, mutations and gene expression have revealed a highly complex assortment of genetic and transcriptional changes within individual tumors. For example, copy number aberrations (CNAs) occur frequently in cancer due to genomic instability. Although multiple new genes have been implicated in cancer through sequencing and CNA analysis[103], these studies have also revealed enormous diversity in genomic aberrations among individuals. Each tumor is unique and typically harbors a large number of genetic lesions, of which only a few drive proliferation and metastasis. Thus, identifying driving mutations (genetic changes that promote cancer progression) and distinguishing them from passengers (those with no selective advantage) has emerged as a major challenge in the genomic characterization of cancer.

The most widely used approaches are based on the frequency an aberration occurs: if a mutation provides a fitness advantage in a given tumor type, its persistence will be favored and it is likely to be found in multiple tumors. For example, GISTIC identifies regions of the genome that are aberrant more often than would be expected by chance, and has been used to analyze a number of cancers[25,187,188]. However, there are limitations to analytical approaches based on CNA data alone: CNA regions are typically large and contain many genes, most of which are passengers that are indistinguishable in copy number from the drivers. CNA data has statistical power to detect only the most frequently recurring drivers above the large number of unrelated chromosomal aberrations that are typical in cancer. Finally, these approaches rarely elucidate the functional importance or physiological impact of the genetic alteration on the tumor. These limitations highlight the need for new approaches that can integrate additional data to identify

drivers of cancer. Gene expression is readily available for many tumors, but how best to combine it with information on CNA is not obvious.

Here we use gene expression as a phenotype, and associate it with CNA. We postulate that driving mutations coincide with a "genomic footprint" in the form of a gene expression signature. We developed an algorithm that integrates chromosomal copy number and gene expression data to find these signatures and identify likely driver genes located in regions that are amplified or deleted in tumors. Each potential driver gene is altered in some, but not all tumors and, when altered, is considered likely to play a contributing role in tumorgenesis. Unique to our approach, each driver is associated with a gene module, which is assumed to be altered by the driver. We sometimes gain insight into the likely role of a candidate driver, based on the annotation of the genes in the associated module.

We demonstrate the utility of our method using a dataset[187] that includes paired measurements of gene expression and copy number from 62 melanoma samples. Our analysis correctly identified known drivers of melanoma and connected them to many of their targets and biological functions. In addition, it predicted novel melanoma tumor dependencies, two of which, *TBC1D16* and *RAB27A,* were confirmed experimentally. Both of these genes are involved in the regulation of vesicular trafficking, which highlights this process as important for proliferation in melanoma.

Our results show that gene expression reflects cellular state and can be used to support algorithms that require information regarding pathway activity and phenotypic outcome. By using gene expression we are able to identify phenotypes and correlate them with locus aberrations. Moreover, we found that many of the phenotypes, and therefore the influence of a driver mutation, are context-specific – present only in a context of another mutation. These results emphasize the critical information gene expression provide over the frequency of a mutation regarding the impact of the mutation, as many driver mutations are present in only a fraction of the tumors.

CONEXIC – a computational framework

We define a "driving mutation" to be a genetic alteration that provides the tumor cell with a growth advantage during carcinogenesis or tumor progression[98]. We reasoned that driving mutations might leave a genomic 'footprint' that can assist in distinguishing between driver and passenger mutations based on the following assumptions:

1. A driving mutation should occur in multiple tumors more often than would be expected by chance (Figure III-1A).
2. A driving mutation may be associated (correlated) with the expression of a group of genes that form a 'module' (Figure III-1B).
3. Copy Number Aberrations often influence the expression of genes in the module via changes in expression of the driver (Figure III-1C).

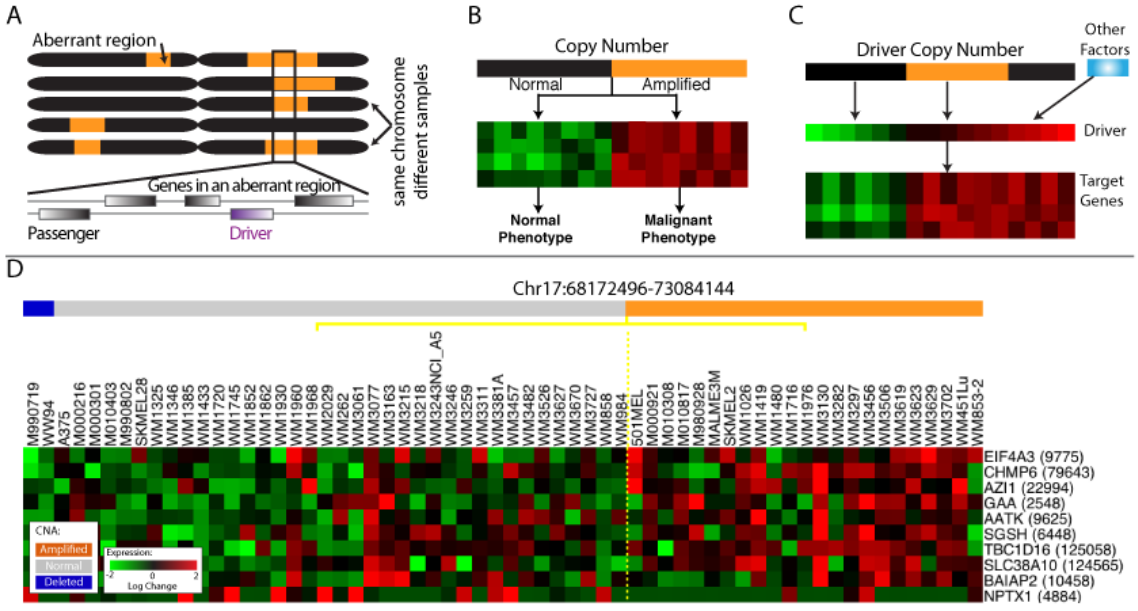


Figure III-1 - The assumptions underlying CONEXIC
For all heat maps, each row represents a gene and each column represents a tumor sample. **A.** The same chromosome in different tumors, orange represents amplified regions. The box shows regions amplified in multiple tumors. **B.** An idealized signature in which the target genes are up-regulated (red) when the DNA encoding the driver is amplified (orange). **C.** A driver may be overexpressed due to amplification of the DNA encoding it, or due to the action of other factors. The target genes correlate with driver gene expression (middle row), rather than driver copy number (top row). **D**. Data representing amplified region on chromosome 17. Heat maps of expression for 10/24 genes in the region that passed initial expression filtering. Samples are ordered according to amplification status of the region (Orange amplified, Blue deleted). These genes are identical in their amplification status and while gene expression is correlated with amplification status to some degree, the expression of each gene is unique. It is these differences that facilitate the identification of the driver.

Driving mutations are frequently associated with the abnormal regulation of processes such as proliferation, differentiation, motility and invasion. Given that many cancer phenotypes are reflected in coordinated differences in the expression of multiple genes (a module)[189,190], a driving mutation might be associated with a characteristic gene expression signature or other phenotypic output representing a group of genes whose expression is *modulated* by the driver. Additionally, CNAs do not typically alter the coding sequence of the driver and so are expected to influence cellular phenotype via changes in the driver's expression. In consequence, changes in expression of the driver are important and so approaches that measure association between the expression of a candidate driver (as opposed to its copy number) and that of the genes in the corresponding module are likely to promote the identification of drivers.

Gene expression is particularly useful for identifying candidate drivers within large amplified or deleted regions of a chromosome: whereas genes located in a region of genomic copy gain/loss are indistinguishable in copy number, expression permits the ranking of genes based on how well they correspond with the phenotype (Figure III-1D). CNA data aids in determining the direction of influence, which cannot be derived based on correlation in gene expression alone. This permits an unbiased approach for identifying candidate drivers from any functional family, beyond transcription factors or signaling proteins.

A Bayesian Network Based Algorithm to Identify Driver Genes

We developed a computational algorithm, COpy Number and EXpression In Cancer (CONEXIC), that integrates matched copy number (amplifications and deletions) and gene expression data from tumor samples to identify driving mutations and the processes they influence. CONEXIC is inspired by Module Networks[169], but has been augmented by a number of critical modifications that make it suitable for identifying drivers (see Computational Methods). CONEXIC uses a score-guided search to identify the combination of modulators that best explains the behavior of a gene expression module across tumor samples and searches for those with the highest score within the amplified or deleted region (see Computational Methods).

The resulting output is a ranked list of high scoring modulators that both correlate with differences in gene expression modules across samples and are located in amplified or deleted

regions in a significant number of these samples. The fact that the modulators are amplified or deleted indicates that they are likely to control the expression of the genes in the corresponding modules (see Figure III-3). Since the modulators are amplified or deleted in a significant number of tumors, it is reasonable to assume that expression of the modulator (altered by copy number) contributes a fitness advantage to the tumor. Therefore, the modulators likely include genes whose alteration provides a fitness advantage to the tumor.

Computational Methods

CONEXIC is a data driven algorithm that takes matched copy number and gene expression data from tumors as input and combines these to identify driving aberrations and associate these with the genes they modulate. CONEXIC is based on a Bayesian scoring function that evaluates each candidate driver, or 'modulator'. The score measures how well a modulator (or combination of modulators) predicts the behavior of a gene expression module across tumor samples. CONEXIC identifies the most likely drivers by searching for the highest scoring modulators in a

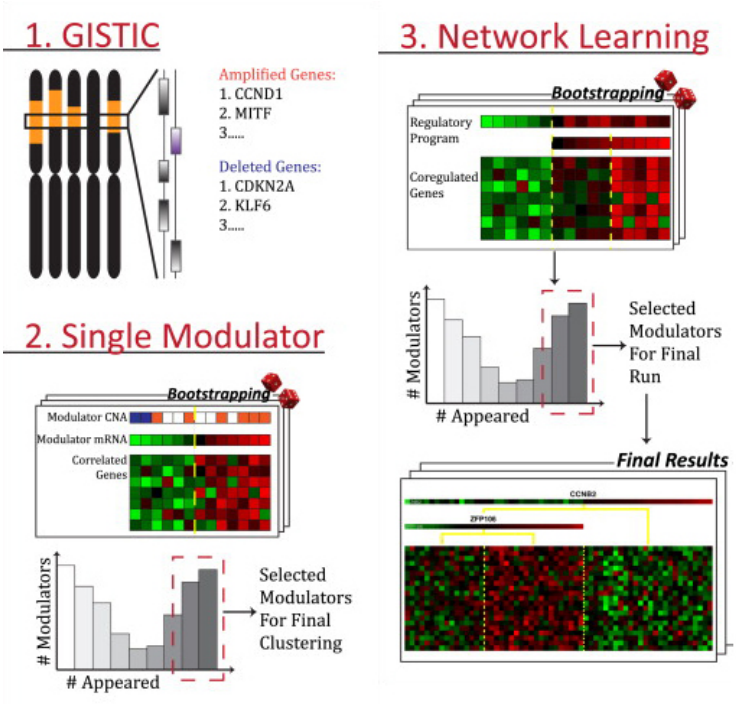


Figure III-2 - Overview of the CONEXIC learning algorithm
1. Selection of candidate driver genes (modulators). The same chromosome is represented in different tumors and orange represents amplified regions. The box shows a region amplified in multiple tumors, considered significantly amplified. Multiple genes reside in this region, represented as shaded boxes along two strands of DNA. Only the purple box represents a driver gene, whereas the gray boxes represent passenger genes and these are indistinguishable based on copy number alone. All genes located in significantly amplified or deleted regions are selected as candidate driver genes. **2.** Single Modulator step. Modules of genes are each associated with the best possible candidate driver, based on gene expression of the gene and the candidate driver. The heat-map represents a good association where both copy number of the modulator influences the expression of the modulator and the expression of the modulator corresponds to the expression of the genes in the module. Random re-sampling with replacement is used to generate perturbations on the initial dataset and this step is repeated across datasets (bootstrapping). The histogram represents the number of runs (datasets) in which each selected modulator appears, the final run is performed with candidate drivers chosen for 90% of the runs (dotted red box). **3.** Network Learning step. Using the set of Single Modulator of modules as a starting point, the algorithm refines the selected modulators and modules, now allowing for more than one modulator associated with each module. Bootstrapping is used similarly to the Single Modulator step, with the distinction being that modulators must be selected in 40% of the runs

stepwise manner (Figure III-2) that seeks to improve the Bayesian score.

The resulting output is a set of predicted modulators that map within an amplified or deleted region. Each is associated with a module, a set of genes whose expression is likely altered by changes in expression of the modulator. In some cases, a module is associated with multiple modulators that form a regulatory program.

The tool is available from http://www.c2b2.columbia.edu/danapeerlab/html/software.html

Regulation programs

We adopted the concept of the regulation program from Segal et al.[169]. Formally, the regulatory program is a conditional probability distribution for the module gene expression, conditioned on the gene expression of the modulators. A regulation program of a module *M* specifies a set of contexts and the expected expression values for each context. A context is determined by the expression of a small set of modulators that influence *M*'s expression. This set of contexts is organized as a regression tree composed of two basic building blocks: decision nodes and leaf nodes. Each decision node corresponds to a modulator and a query on its value (for example, "is MITF >= threshold"). Each decision node has two child nodes: the right child node is chosen when the answer to the corresponding query is true; the left node is chosen when it is false. For each sample, one begins at the root node and continues down the tree in a path according to the answers to the queries in that particular sample until a leaf node is reached. Each leaf encodes a probability distribution representing how the module's genes are expected to behave in that sample. The expression of genes in *M* in each context is modeled as a normal distribution; this distribution is encoded using a mean and variance stored at the corresponding leaf.

Regression trees are particularly well suited for modeling driving mutations in cancer because (i) these can capture combinatorial and condition specific relations that frequently occur in cancer (e.g. both over-expression of EGFR and deletion of P16 are required) (ii) These can capture changes in both the mean and the variance of the module gene expression.

The CONEXIC algorithm

The CONEXIC learning algorithm consists of three key steps:

1. Selection of candidate driver genes (Figure III-2A)
2. Single Modulator step that creates an initial association between candidate drivers and gene modules (Figure III-2B)
3. An iterative Network Learning step to improve on the initial model (Figure III-2C)

CONEXIC searches for a model that can explain the variation in the gene expression across samples as a function of a small number of modulators. The search is driven by the optimization of a Bayesian scoring function. The search begins with an initial starting point and then proceeds in making stepwise changes that improve the score at each step.

Selection of Candidate Drivers

Motivation: First, we want to identify regions of the DNA that are recurrently amplified/deleted in tumors and consider genes within or neighboring to those regions as candidate drivers. We expect that many of the driving mutations will be contained in this candidate list.

Details: We applied the (GISTIC) algorithm[25], using the JISTIC implementation available from http://www.c2b2.columbia.edu/danapeerlab/html/software.html[104], to all 101 samples. To increase sensitivity, we used a q-value threshold of 0.3, compared to 0.25 previously used with this data [187]. Genes that overlap a significant aberrant region are chosen as candidate driver genes. To capture aberrations in regulatory regions, for each aberrant region, the closest non-overlapping genes on each side are also chosen as candidate drivers, if their distance from the edge of the region is less than 100Kb.

Result: This resulted in 27 amplified regions containing 513 peak genes and 23 deleted regions with 384 peak genes (Table III-1). In subsequent steps, we aim to identify which of these 897 candidates are likely drivers.

Expression Filtering: We now integrate copy number and gene expression data, which is available for 62 tumors. As an initial filter, we require candidate drivers to be differentially

expressed across the different tumor samples. This removes genes that are expressed at a constant level across all tumors and not influenced by their copy number. Additionally, this removes genes that are not expressed, and are therefore unlikely to be drivers. Our final set of candidate drivers included 428 genes in significantly amplified or deleted regions, whose gene expression varied with standard deviation greater than 0.25.

Single Modulator step

Motivation: The Single Modulator step constructs an initial model, which focuses the subsequent search on variation in gene expression that can be explained by drivers encoded in CNAs (as opposed to variation due to other types of aberrations such as coding mutations). The Single Modulator step (Figure III-2B), establishes an initial pairing between candidate drivers and gene expression modules by associating each target gene with the single driver gene that fits it best (based on corresponding gene expression profiles). As a result, each gene is clustered into a module consisting of those genes for which the same driver gene was found to be the best fit and the module is associated with that candidate driver.

Details: To aid the identification of a good starting model, the Single Modulator step considers a smaller search space and more conservative set of candidate drivers, only those whose gene expression is significantly altered by either their amplification or deletion status. Candidates are filtered using a Welch t-test (p-value < 0.05), comparing amplified versus normal or deleted versus normal - 347 candidate drivers pass this test. Amplification or deletion status of a gene in a specific sample is determined using the average copy-number value for all SNP markers inside the gene; if the gene contains no SNP markers we take the copy-number value for the single closest SNP marker. Using the same thresholds as in [187], if the average value of the SNP markers around a gene is above 0.3, the gene is marked as amplified, if the average value is below -0.3,it is marked as deleted.

For each candidate driver gene, we use the gene expression values of the amplified/deleted samples to guide the choice of threshold, and consider the gene expression of the amplified/deleted samples to represent appropriate high/low expression levels. We use k-means clustering, using k=2 and the normal and amplified/deleted samples as the two initial clusters to fit

two normal distributions. The boundary between the two clusters is the selected expression threshold level for this driver gene. The expression of each target gene is split into two sets: those in the tumor samples in which the driver's expression is below the threshold, and those in the tumor samples in which the driver's expression is above the threshold. The NormalGamma scoring function is used to compute the quality of this split, thus measuring a target gene's fit with a candidate driver.

After the score is computed for all pair-wise combinations of candidate drivers and target genes, each gene is assigned to the single highest scoring candidate driver. Permutation testing is used to verify the statistical significance of association between driver and gene. Driver gene expression is randomly permuted 10,000 times, conserving the composition of values, but rendering the order random and independent of the target gene. Each of these permutations is scored, creating a null distribution to compare with the unpermuted order; thus providing a p-value for the association between gene and candidate driver. If this p-value < .001, we associate between gene and candidate driver, declaring the candidate driver a modulator of its associated gene. Care must be taken to avoid spurious associations due to the dense correlation structure of genes encoded in the same aberrant region, as it is easy to obtain associations between all candidate drivers in a region with the same target gene. Thus, it is important to find only the single best association between a gene and its modulator.

We established a number of additional criteria to ensure the robustness of our results; these criteria were guided by results on randomly permuted data. First, we require that each modulator be chosen by at least 20 genes. If some of the modules have less than 20 members, we break up the smallest module and reassign its genes to the next best scoring modulator, repeating until all modules have at least 20 members. Second, we apply non-parametric bootstrapping and repeat this procedure 100 times, generating variations on the dataset using random re-sampling with replacement. This ensures that the association is not an artifact of the specific set of samples and is robust across different subsets of the original data. We select candidate drivers that were selected in at least 90% of the runs.

We then make one final run of the Single Modulator step, using this filtered set of 130 candidate driver genes. The selected set of modulators all reside in significantly aberrant regions and their gene expression best corresponds with at least 20 genes. The statistical significance of this association is ensured both by permutation testing and non-parametric bootstrap.

Result: Single Modulator step identified 78 modulators that explain the behavior of 4018 genes. Each of the 78 modulators is associated with a module containing at least 20 genes. These will be refined in subsequent steps.

Network Learning step

Motivation: The Network Learning step (Figure III-2C) uses the modules generated by the Single Modulator step as a starting point and uses an iterative approach to improve the score of the modules and their regulatory programs. The Network Learning step is based on the Module Networks algorithm[169,174] with a few critical improvements, designed to remove spurious association, described below.

Details: The algorithm iteratively alternates between two tasks: (i) learning the regulation program for each module; (ii) and re-assigning each gene into the module that best models its behavior. The score improves at each iteration and these terminate if fewer than 10% of the target genes have been re-assigned to a different module during the gene re-assignment step.

Given a set of modules, we learn a regulatory program for each module. We recursively learn the regulatory program by choosing, at each point, the candidate driver that best splits the gene expression of the module genes into two distinct behaviors. All candidate drivers and potential split values are evaluated and the driver-split combination that achieves the highest improvement in score is selected. Only if the score improvement is greater than a pre-defined penalty, the split is selected. Unlike Single Modulator, all 428 candidate drivers are considered and each candidate driver is not limited to a single split threshold; rather the optimal threshold is chosen to maximize the score. While multiple thresholds are possible across the different regulatory programs, the score includes a penalty on the number of different split values, thus limiting the number selected. The tree is recursively grown from the root to its leaves. At each new split, the driver gene that provides the best improvement in score is chosen; permutation testing (as in the Single Modulator

step) is used to ensure the statistical significance of the split; and the outlier-removal test (described below) is used to ensure the modulator was not selected due to outliers in the data. For each new split, linear influence of the modulator on each side of the split is tested (as described below), and if linear influence is found further sub-splits on this side of the split are forbidden. The process terminates when no query that improves the score and passes these two tests can be found, allowing for up to a total of five splits in each regulatory program.

In addition to permutation testing, the regulatory-program learning process includes two additional statistical tests, designed to remove spurious splits. (i) To ensure that modulators were not selected due to outliers in the data, splits are subjected to a outlier-removal test, as follows: for each candidate split, we remove the highest 4% of expression values in the side that has higher mean expression, and re-calculate the score improvement with the remaining 96% of the data. We reject the split if the score improvement is less than 0.6 times the score improvement with the entire data. We perform a similar test removing the lowest 4% of expression values in the side that has lower mean expression. The resulting splits are robust to outliers in the data.

(ii) Some of the modulators have a linear influence on the target gene expression and their values are enough to explain the variation in the expression of the target genes without additional splits. In most cases, such influence is only on one side of the split, e.g. target gene expression of one leaf is linearly correlated with the modulator expression, while the expression of the other leaf is not. Our goal is to remove additional splits when the modulator is correlated across all samples and retain splits in cases the correlation is one sided. Correlation alone is not sufficiently sensitive to distinguish between a strong correlation that is limited to one side of the split, versus a weaker correlation across all samples. To make this distinction, for each new split, and for each side of that split, we calculate the Pearson correlation and regression slope between the expression values of the modulator and of the members of the module. First, we ask whether the modulator is correlated with the module, and require a Pearson correlation coefficient > 0.6 for at least one of the leaves. Next, we evaluate whether the slope on both sides of the split is similar, requiring that the slope in the leaf with higher correlation and the slope of the combined data (on both

leaves) be within a ratio of 0.7 to1.4. If both criteria hold, we forbid any further sub-splits on the correlated side.

Given the inferred regulation programs, we determine the module whose associated regulation program best predicts each gene's behavior. Specifically, we iterate over all genes, one at a time, and move each gene into the module that provides the highest improvement in the score. This step is guaranteed to improve the score, or leave it the same (if the gene is not moved). We repeat this reassignment process for all genes three times, at every iteration.

Similarly to the Single Modulator step we boost robustness using non-parametric bootstrap. The iterative learning algorithm is run 100 times. We then filter the set of candidate driver genes, leaving only genes that appeared in at least one regulatory program in at least 40% of the runs. 65 modulators pass this threshold and continue to the next step. We then make one final run of the Network Learning algorithm, using this filtered set of candidate driver genes.

Results: This resulted in the identification of 64 modulators that explain the behavior of 7869 genes. We compared the models at the beginning versus at the end of the Network Learning step and found the end model superior by a number of measures:

1. The final model can explain the behavior of 7869 genes, relative to only 4018 at the end of Single Modulator (starting model).
2. The test log-likelihood is significantly higher for the final model, relative to the initial model, in a leave-one-out cross validation (see Figure III-2D and robustness section below).
3. *TBC1D16* is a good example of the need for the more aggressive search performed by network learning. It was the 2nd highest scoring modulator at the end of network learning and empirically validated. Due to its more limited search, Single Modulator did not select *TBC1D16* at all.

Model refinement

The candidate drivers used for the regulatory programs include only genes residing in CNA regions, which are expected to explain only part of the global changes in gene expression. Observed changes in gene expression can also result from additional factors, such as somatic

mutations that are not included in our data. While the algorithm attempts to assign all genes to modules, some genes expression is not influenced by any CNA based driver. Therefore, it is important to remove genes that can't be explained by any regulatory program, or more formally, no program significantly improves the gene's likelihood. For each gene, we calculate the difference between the likelihood of its data using its assigned regulatory program and the likelihood of its data without any regulatory program. For each module, we calculate the distribution of these delta likelihoods; those genes for which the delta likelihood is two standard deviations or more below the mean are removed from the module. These genes can't be explained by any regulatory program and will not be members of any module. A final iteration of learning the regulation program is executed after these genes are removed.

Score function

We use a Bayesian scoring approach that maximizes the overall joint probability of both the data and of the model structure. Let D represent the data and S represent the structure of the network, then the scoring function is expressed as $\log P(D,S) = \log P(D \mid S) + \log P(S)$. Where the first term is the likelihood of the data for a given model (in the Bayesian approach we integrate over all possible model parameters) and the second term is the prior on the structure for which we use a penalty score on model complexity.

Following the Module Networks approach[169] we use Normal Gamma distribution for our likelihood function. Normal Gamma gives a higher score to data with lower variance and hence finds splits that create two different contexts that represent two distinct behaviors (normal distributions). The Normal Gamma score is described below:

$$NormalGamma(Leaf,\lambda,\alpha):$$

$$N = Size(Leaf)$$

$$\beta = Max(1,\frac{\lambda*(\alpha-2)}{\lambda+1})$$

$$\beta^{+} = \beta + \frac{Var(Leaf)*N}{2} + \frac{N*\lambda*\overline{Leaf}^{2}}{2*(N+\lambda)}$$

$$\alpha^{+} = \alpha + \frac{N}{2}$$

$$Score = -N*\ln(\sqrt{2\pi}) + \frac{\ln(\frac{\lambda}{\lambda+N})}{2} + \ln(\Gamma(\alpha^{+})) - \ln(\Gamma(\alpha)) + \alpha*\ln(\beta) - \alpha^{+}*\ln(\beta^{+})$$

Leaf is a vector of gene expression values contained in the leaf and α,λ and β are parameters. A split is scored by comparing the score of the split data to the score without the split, along with a penalty for the split.

$$NormalGamma(Left\_Leaf) + NormalGamma(Right\_Leaf) > NormalGamma(Entire\_data) + Penalty$$

Our penalty function is comprised of two parts. Following Module Networks [169], we use a complexity prior that penalized the number of leaves in each regulation program, using the exponential distribution over total number of leaves. Denoting the regulatory program as T and L as number of leaves, $\log P(T) = -\beta L$. Following genetic Module Networks [174], in addition to a penalty specific to each regulation program, we have a network wide penalty function that penalizes the total number of modulators. The prior takes the form of a power-law distribution on the number of modulators. This prior encourages the algorithm to select a sparse number of modulators, which is particularly important in this application, whose main purpose is to identify a small set of potential drivers. Full details are available in[174].

The scoring function has 5 parameters, α and λ for the Normal Gamma distribution and β, x and y for the complexity prior. These were selected using 10-fold cross validation and the parameters used were α=2, λ=1, β=20, x=15 and y=0.

Parameter Selection and Robustness

Selection of Candidate Drivers

Selection of candidate drivers requires determining a q-value threshold for GISTIC, the higher the threshold, the more candidate regions and genes will be selected, 0.25 is typically used as a threshold for determining the final list of significant regions [25,187,188]. Within CONEXIC, GISTIC is used to only generate a pool of candidate genes for further selection, so we used the more permissive threshold of 0.3. It is likely that there are additional drivers even beyond a threshold of 0.3, but too many candidate modulators burden CONEXIC both computationally and statistically. Therefore, we selected a threshold of 0.3 and correctly identified *CCNB2* and *RAB27A* as drivers in region below the 0.25 threshold, demonstrating increased sensitivity.

Single Modulator step

The Single Modulator requires a confidence threshold for non-parametric bootstrap. We selected 90, meaning that we only selected modulators chosen in more than 90% of the bootstrap runs. Before removing modules containing fewer than 20 genes the median single modulator run included 295 modulators. After removing small modules, a median of 202 modulators still remained. Following bootstrap with a threshold of 90% only 78 remained.

Why did we choose 90? In a histogram representing the number of modulators at each confidence threshold (Figure III-3A) we observe that below 90 the distribution of modulators at each confidence level flattens and becomes uniform. It is important to note that this threshold does not define a filter, but rather only a starting point for Network Learning, which reconsiders all 428 candidate drivers. Indeed, 10 modulators that are not selected at this stage are included in the final model, including *TBC1D16* and *ZPF106*.

CONEXIC achieves similar results across a broad range of thresholds and the final results bear significant similarity, even in a comparison between using 80 versus 95 as a threshold. Using 80 as a threshold results in 60 modulators and using 95 as a threshold results in 57 modulators, the overlap between them is 45 modulators. The final model using 80 as a threshold

is much closer to our final model, with complete overlap in all the modulators discussed in this chapter.

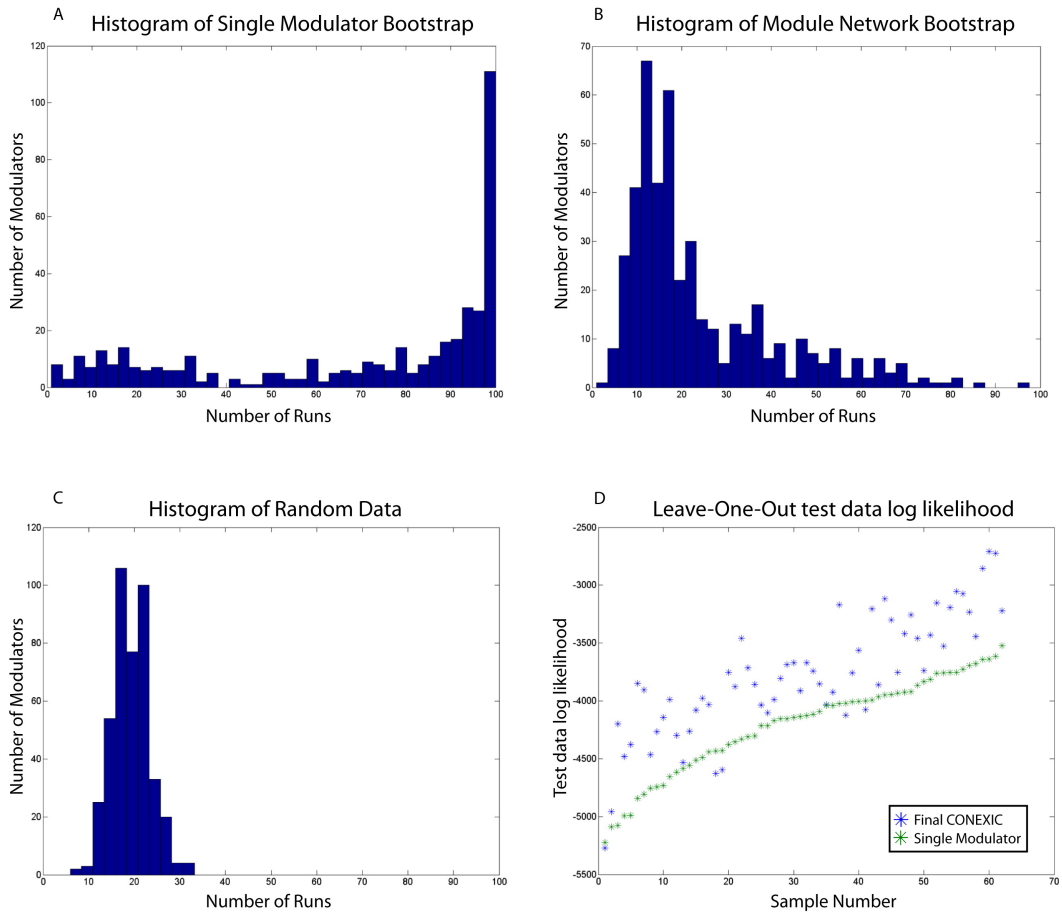


Figure III-3 - Robustness analysis

A. Histogram representing confidence values for bootstrap of Single Modulator. **B.** Histogram representing confidence values for bootstrap of Network Learning. **C.** Histogram representing confidence values for bootstrap of Network Learning, where the data has been randomly permuted. These values represent spurious modulators and none passed a confidence threshold of 35. **D.** Test likelihood for each sample in leave-one-out cross-validation. Each column is matched for the same sample, where green and blue stars represent test likelihood for single modulator and test likelihood for final CONEXIC model, respectively. The full model significantly outperforms Single Modulator on test data.

Network Learning step:

While the model resulting from Network Learning is higher scoring than the model resulting from Single Modulator, the former has more parameters and hence over-fitting is a concern. To evaluate whether the learned models can be generalized to unseen data, we compared the models derived from Single Modulator versus Network Learning using leave-one-out cross validation. For each sample (tumor), the two models were learned using the 61 other samples.

We calculated the test log-likelihood of the held out 'test' sample for each of the models. In essence, we are testing how well the model "predicts" a new sample. The likelihood of the held out sample was consistently better in the Network Learning model, for almost all of the 62 samples (Figure III-3D).

Network learning reconsiders all candidate drivers, so bootstrap is needed for this step, for the same reason it is needed in Single Modulator. Here, a clear threshold cannot readily be determined from a histogram representing the number of modulators at each confidence threshold (Figure III-3B). Instead, we used a similar histogram generated from the same data, only randomly permuted to determine a confidence threshold that is beyond spurious associations. Observing the histogram in Figure III-3C we see there are no spurious associations above 35, so we selected 40 as our threshold.

A concern in this domain is its dense correlation structure; candidate drivers residing in the same region are all correlated to the region's copy number and among themselves. Bootstrapping has a pivotal role in dealing with the spurious correlations that can arise: if a candidate driver corresponds to a module only through its correlation to its own copy number, other candidates in the region are equally likely to be selected and will not hold pass bootstrapping. Only when the candidate driver provides a substantially better score than its neighboring candidates in the region does the association hold. A threshold of 40 is more than enough to serve this role, as most regions have at least 5 genes, typically dozens.

This threshold largely determines the final output and changes in it directly add or remove modulators from the final model. The result is a relatively short, ranked list of putative drivers, however discretion must still be applied. The ranking is informative; the top 10 are more reliable than those ranked 50 to 60. Additional filters can be applied to evaluate this list, such as whether it replicates in other datasets or whether it associates with a module annotated for cancer related processes. Our goal was generating a list of modulators that have a signal above spurious noise using a less stringent threshold, leaving the final evaluation to the informed user.

Comparison to other methods

Comparison to Module Networks

CONEXIC uses an algorithm similar to Module Networks[169] as its key statistical engine. Module Networks is based on two principles: (i) influences and interactions between proteins often generate statistical dependencies in gene expression and (ii) testing dependencies on entire modules of genes enables statistical discovery that is undetectable when considering each gene in isolation. Module Networks was designed to infer regulatory models of transcription factors and their upstream signaling proteins. Module Networks handles the ambiguity between correlation and influence using prior knowledge: Taking a precompiled list of transcription factors and signaling proteins it assumes: 'If a protein that has a known role in transcriptional regulation and is correlated (or anti-correlated) with the expression of genes in a module, it is likely to regulate the genes in that module.' Module Networks was applied to a yeast dataset with 173 samples, to build a regulatory model over 2355 genes[169].

CONEXIC has a fundamentally different goal: the identification of drivers of cancer. In this application, the primary role of the module is to provide support for a gene as a driver, the 'network interpretation' is secondary. We apply CONEXIC to a human cancer dataset with 62 samples, to build a model over 7869 genes involving influences that go beyond direct transcriptional cascades. CONEXIC is based on a different set of assumptions. It assumes that perturbations originate in the DNA and this provides the direction of influence (Figure III-3A). The set of candidate modulators includes all genes contained within frequently amplified and deleted regions of any functional class and thus extends beyond transcriptional regulation. It aims to capture modulation of expression in response to altered cell physiology (Figure III-3B). Aberrations in the DNA lead to perturbations which provide a rich source of variation that can be used to help uncover molecular influences in the cell. Numerous adaptations were made to the Module Networks algorithm to make it more suitable for this application.

We provided the Module Networks algorithm, as originally implemented[169] with the same candidate driver set (GISTIC output) used by CONEXIC. The resulting model includes 347 modulators, which are a large fraction of the 428 candidates that pass expression filtering and is

of limited use as a selective method for identifying drivers. Additionally, only one of the empirically confirmed MITF targets (as defined by Hoek (Hoek et al., 2008) was associated with MITF, in contrast to the 45 identified by CONEXIC. The Single Modulator step was used to derive modules for initialization of Module Networks, as opposed to clustering initialization. Starting from the Single Modulator (with bootstrap) defined modules, the final output included 109 modulators, and MITF was associated with 46 of its experimentally derived targets (both numbers are median values across 100 runs). We conclude that the Single Modulator initialization method is critical for CONEXIC's success in associating a modulator to genes it influences. It focuses the learning on changes in expression altered by CNA, as opposed to those altered via other mechanisms (e.g. coding mutations in the ERK pathway).

The output from running Module Networks with bootstrap, using a candidate set defined by GISTIC and initialized using Single Modulator, yields results similar to those of CONEXIC. Each of the additional refinements: permutation testing, outlier removal, removal of linear splits and gene removal provide smaller improvements to the final results. For example, the removal of linear splits only removes 21 out of 489 splits in the model. Nevertheless, each of these steps improves the model, as assessed by leave-one-out likelihood tests.

Integration of CNA with gene expression

Methods based on copy number information alone, e.g. GISTIC[25,188] are typically limited to detecting large regions containing multiple genes, such that the driver cannot be readily identified among them. Applied to this melanoma dataset, GISTIC (with q-value of 0.25) finds multiple regions containing 588 genes[187], the drivers and passengers indistinguishable. While GISTIC is a valuable method to filter a genome of ~23,000 genes and derive a set of hundreds of candidate drivers, additional data types are required to narrow this list down further.

A number of different approaches integrate CNA and gene expression by identifying genes with significant correlations between DNA copy number and gene expression. Lin et al.[187] applied the approach to this data and predicted KLF6 and CUL2 as putative drivers. Recently Huh et al.[191] KLF6 was validated as a driver in melanoma. MITF CNA is poorly correlated with its gene expression and hence not identified with this approach.

SLAMS[192] bears some conceptual similarity to CONEXIC, but there are critical differences. SLAMS requires an initial signature that is used to divide the samples into classes and runs SAM to find the copy number that best separates the classes. The algorithm then finds a gene (or multiple genes) in the selected region, where the expression of the gene is a good predictor of the expression signature. In contrast to SLAMS, CONEXIC does not require a pre-defined expression signature, but identifies one or more signatures de novo. To test SLAMS on the melanoma dataset, we used the MITF targets identified by Hoek[193] as a signature. SLAMS identified the copy number of 1444 genes as significant, ranking MITF as 75th. In contrast, CONEXIC correctly identified MITF as the top ranked gene, associated MITF with its targets de-novo and predicted many additional drivers.

Witten et al.[194] described a method based on applying penalized canonical correlation analysis (CCA) to the cross product matrix of gene expression and CNA data, identifying the regions and correlated genes in a single step. We applied the method to the melanoma dataset using the same steps and parameters as those used in the original paper. This method identified 7980 genes as significant, including almost all the genes influenced by CNA, but did not distinguish the drivers among them.

Methods that integrate CNAs with other data types

In addition to expression, other data types have been used with CNA to help identify drivers. GRAIL[195] prioritizes genes within GISTIC regions based on prior knowledge and known gene annotation. GRAIL identified MCL1 (using 3000 samples across multiple cancers), but failed to find MITF or KLF6[61].

Another approach, NetBox[196], uses a curated human protein-protein interaction database as an additional source of information. It constructs protein-protein interaction networks from genes within recurrently aberrant regions and defines drivers as hubs in these networks. Applied to the Lin melanoma dataset NetBox did not find any significant networks (lowest p-value 0.15). Even considering networks of low significance, NetBox did not identify MITF, KLF6 or any other known melanoma oncogene/tumor suppressor. Both GRAIL and NetBox are strongly based on prior

knowledge and annotations. While they present a powerful approach for identifying oncogenes in new contexts (e.g. MCL1 which has not yet been verified in melanoma), they only predict drivers among well annotated genes. The advantage of assaying copy number, gene expression, sequencing and other technologies genome-wide is that the data are comprehensive and unbiased. To fully exploit this data we need methods that go beyond the realm of the well annotated.

Results – CONEXIC in melanoma

We applied the CONEXIC algorithm to paired gene expression and CNA data from 62 cultured (long and short term) melanomas[187]. A list of frequently altered loci was generated using copy number data available for 101 melanoma samples by applying a modified version[104] of

Gene Symbol	Pathway	Band	Genes in Region	Validation p-value
MITF	Melanoma	3p14.2-p14.1	1	$<10^{-6}$
TBC1D16	Vesicular Trafficking	17q25.3	24	$<10^{-6}$
ZFP106	Insulin/Ras	15q15.1	7	$<10^{-6}$
DIXDC1	Wnt/JNK/PI3K	11q23.1	17	0.0001
OIP5	Cell Cycle	15q15.1	13	$<10^{-6}$
TTBK2		15q15.2	7	0.0383
TRAF3	NFkappaB/JNK	14q32.32	19	0.0121
RAB27A	Vesicular Trafficking	15q15-q21.1	33	$<10^{-6}$
C12orf35		12p11.21	45	$<10^{-6}$
WBP2		17q25	92	0.0275
MOCS3		20q13.13	16	$<10^{-6}$
NDUFB2		7q34	10	$<10^{-6}$
ST6GALNAC2		17q25.1	92	$<10^{-6}$
GRB2	EGFR/Ras	17q24-q25	92	0.1373
ECM1		1q21	55	0.0083
KCNG1		20q13	16	0.202
DPM1		20q13.13	16	0.097
PFKP	Metabolism	10p15.3-p15.2	3	0.0801
KLF6	Cell cycle, c-JUN (JNK)	10p15	3	$<10^{-6}$
TIMM8B	Mitochondria	11q23.1-q23.2	17	0.7622
PI4KB		1q21	55	0.0003
PSMB4		1q21	55	0.0005
VPS72		1q21	55	$<10^{-6}$
TARS2		1q21.3	55	0.0001
MNS1		15q21.3	33	0.0908
TDRD3	RNA processing	13q21.2	203	$<10^{-6}$
CCNB2	Cell Cycle	15q22.2	33	$<10^{-6}$
EIF5	Cell Cycle	14q32.32	19	0.1096
RAB7A	Vesicular Trafficking	3q21.3	16	$<10^{-6}$
PIK3CB	PI3K signaling	3q22.3	15	$<10^{-6}$

Melanoma | Oncogene or Tumor Suppressor

Figure III-4 - Top 30 modulators
Gene names are color-coded based on the role of the gene in cancer, 10 genes have been previously identified as oncogenes or tumor suppressors (peach), of these 3 in melanoma (brown). Column 3 represents chromosomal location, where orange represents amplification and blue represents deletion. These genes were identified within regions containing multiple genes, the number of genes in each aberrant region is listed in column 4. Column 5 lists the p-value for modulator validation in independent data. p-values are shown for the Johansson dataset, unless the modulator was missing from this dataset, and then p-value from the Hoek dataset is shown.

GISTIC[25] (see Table III-1). Next, we integrated copy number and gene expression data (available for 62 tumors) to identify the most likely drivers within those loci (see Computational methods). Statistical power is gained by integrating all data, and by combining statistical tests on thousands of genes to support the selected modulators. This resulted in the identification of 64 modulators that explain the behavior of 7869 genes. We consider the top 30 scoring modulators, presented in Figure III-4, as likely drivers.

Table III-1 - List of aberrant regions

Amplifications				Deletions			
Chr.	**Start**	**End**	**#Genes**	**Chr**	**Start**	**End**	**#Genes**
1	147226092	150596000	55	5	58445032	58683084	1
3	70107048	70229264	1	5	59075480	59432304	1
3	120500072	120923376	12	5	112049568	11210581	1
3	129929104	130624856	16	6	162494042	16363721	2
3	139476272	139507824	4	8	6333250	9359366	28
3	139674960	141259840	15	9	21999960	22009732	2
3	173931152	174312880	3	10	89436437	89908984	3
3	175468448	175633120	1	11	111213008	11196167	17
3	192266400	199124224	52	13	19867988	96252808	203
5	11347004	14472551	3	14	38302632	38988776	7
6	3859295	5698904	9	14	45505796	46786096	2
6	57103452	57277264	5	14	102319430	10381078	19
7	139232720	140249792	10	15	39444436	39948236	13
8	121760777	128993129	31	15	40049072	40114644	4
1	68953208	69754234	8	15	40603172	40933120	7
1	18722300	19574988	5	15	41605345	43473384	30
1	24388521	32822550	45	16	52162032	52621120	3
1	67491552	67636136	3	16	77264113	78880878	2
1	53109188	58337128	33	16	87889112	87959104	2
1	87547925	89038234	23	18	62719491	76117153	33
1	68172496	73084144	92	7	129619320	13033013	6
2	47893352	49179608	16	7	12105143	13052351	2
2	39399572	40948612	36	17	74741058	77061186	24
				20	49142876	50443163	4
				10	2552329	4073842	3
				14	55290464	55652536	2
				14	57178720	57588888	2

This table specifies the significant regions of copy number gain (A) and copy number loss (B) and the correlation between copy number and expression for all candidate regulators

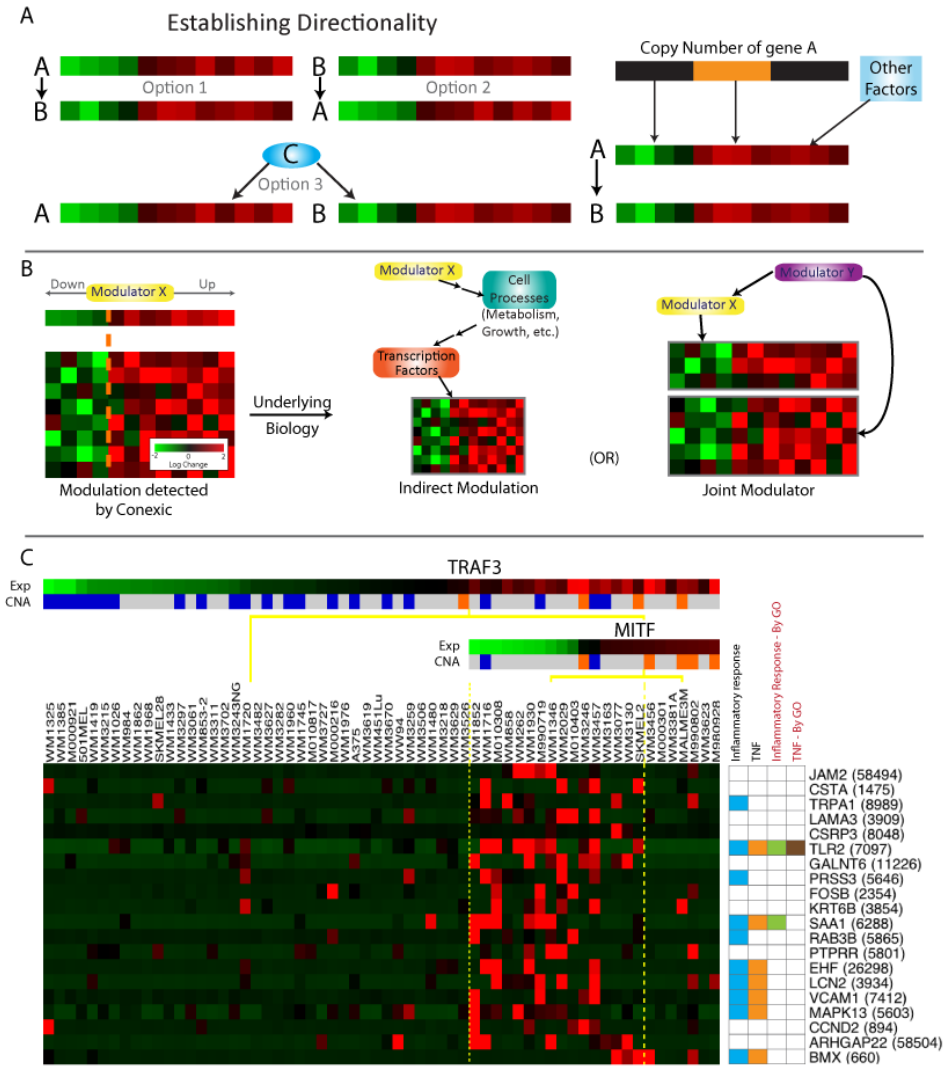


Figure III-5 - Associating modulators to genes

A. Three scenarios could explain a correlation between a candidate driver (gene A) and its target (gene B): A could influence B, B influence A, or both could be regulated by a common third mechanism. The availability of both gene expression and chromosomal copy number data allows us to establish the likely direction of influence. If the expression of gene A is correlated with its DNA copy number, and the copy number is altered in a large number of tumors, it is likely that the copy number alteration results in a change in expression of A in these tumors. So the model in which A influences the expression of B and other correlated genes is the most likely. In this way, examination of both copy number and gene expression in a single integrated computational framework facilitates identification of candidate drivers. **B.** Modulator influence on a module can go beyond direct transcriptional cascades involving transcription factors or signaling proteins and their targets. Genetic alteration of any gene (e.g., a metabolic enzyme) can alter cell physiology, which is sensed by the cell and subsequently leads to a transcriptional response through a cascade of indirect influences and mechanisms. While modules are typically enriched for genes influenced by the modulator, they also contain genes that are coexpressed with the modulator ('joint modulator'). Both types are helpful for annotating the module and determining the functional role of the modulator. **C.** The TNF module. The modulators include *TRAF3* and *MITF*, where high *TRAF3* and low *MITF* are required for upregulation of the genes in the module. The annotation for each gene is represented in a color-coded matrix. Blue and orange squares represent literature-based annotation; green and brown are from GO. LitVAN associated the genes in this module with TNF and the inflammatory response.

Many modulators are involved in pathways related to melanoma

The top 30 modulators (likely drivers) include 10 known oncogenes and tumor suppressors (Figure III-4). In many cases, CONEXIC chose the cancer related gene out of a large aberrant region containing many genes. For example, *DIXDC1*, a gene known to be involved in the induction of colon cancer[197], was selected among 17 genes in an aberrant region. *CCNB2*, a cell cycle regulator, was selected from a large amplified region containing 33 genes. The modulators span diverse functional classes including: signal transducers (*TRAF3*), transcription factors (*KLF6*), translation factors (*EIF5*) and genes involved in vesicular trafficking (*RAB27A*).

Performing a comprehensive literature search for all genes is tedious and time consuming, so we developed an automated procedure, LitVAn - Literature Vector Analysis, that searches for over-represented terms in papers associated with genes in a gene set. LitVAn uses a manually curated database (NCBI Gene) to connect genes with terms from the complete text of more than 70,000 published scientific articles (see LitVAn). LitVAN found a number of over represented terms (Figure III-11) among the top 30 modulators, including 'PI3K' and 'MAPK', which are known to be activated in melanoma, 'cyclin', representing proliferation which is common in all cancers and 'RAB'. Rabs regulate vesicular trafficking, a process not previously implicated in melanoma[19].

The association between a modulator and the genes in a module

Beyond generating a list of likely drivers (modulators), the CONEXIC output includes groups of genes that are associated with each modulator (modules). We tested how reproducible the modulators and their associated modules are using gene expression data from two other melanoma cohorts with 45[198] and 63[199] samples (see Comparison to other methods). We found that 51/64 (80%) of the selected modulators are conserved across datasets in a statistically significant manner. Modules (statistically associated genes) are likely enriched with genes whose expression is biologically affected by the modulator (Figure III-5). In consequence, the processes and pathways represented by genes in a module can help us to gain insight into how an aberration in the modulator might alter the cellular physiology and contribute to the malignant phenotype.

Annotation of data-derived sets of genes is typically carried out based on gene set enrichment using Gene Ontology (GO) annotation. Although this approach is useful, there are modules for which GO annotation does not capture the known biology. For example, the 'TNF module' is enriched with the GO terms 'developmental process' and 'cell differentiation' (q-value=0.0014 and 0.004 respectively). We used LitVAn to carry out a systematic literature search and found 11/20 genes in the module related to the TNF pathway, inflammation or both (Figure III-5C), although only 2 of these genes were annotated for these processes in GO. *TRAF3*, the modulator chosen by CONEXIC, is known to regulate the NF-kappa-B pathway[200], a major downstream target of TNF. Although *TRAF3* has not been previously implicated in melanoma, the importance of the NF-kappa-B pathway in melanoma is well supported.

A known driver, MITF, is correctly associated with target genes

CONEXIC identified microphthalmia-associated transcription factor (MITF) as the highest scoring modulator. MITF is a master regulator of melanocyte development, function and survival[119,201] and the over-expression of MITF is known to have an adverse effect on patient survival[103].

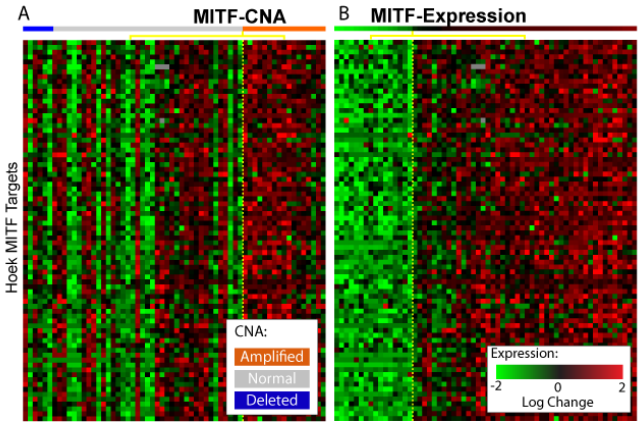


Figure III-6 - *MITF* expression correlates with expression of the genes in the associated module
A. Each row represents the gene expression of one of 78 MITF targets identified by Hoek [193]; the tumor samples are split into two groups based on the copy number of *MITF* (Welch t-test p-value=0.04) **B.** The rows represent the same genes, in the same order as in A, but here the tumor samples are split into a group of samples that express *MITF* at high (n=46) or low levels (n=16) Welch t-test p-value=0.0001).

To test the association between modulator and module, we obtained an experimentally derived list of MITF targets [193] and asked whether the modules identified by CONEXIC associate MITF with its known targets. The *MITF* associated modules contained 45/80 previously identified targets (p-value < 10^{-45}) supporting a match between the transcription factor (TF) and its known targets. However, a few targets (*TBC1D16*, *ZFP106* and *RAB27A*) are both associated with MITF and are themselves

modulators of additional modules. CONEXIC limits each gene to a single module, so association with an MITF target would preclude association with MITF. If we permit indirect association to MITF through the modules of these additional modulators, CONEXIC correctly identifies 76 of the 80 targets identified by Hoek *et al.* (p-value<10^{-78}). Similar target sets are not available for any other modulator, precluding a more rigorous evaluation of our other predictions.

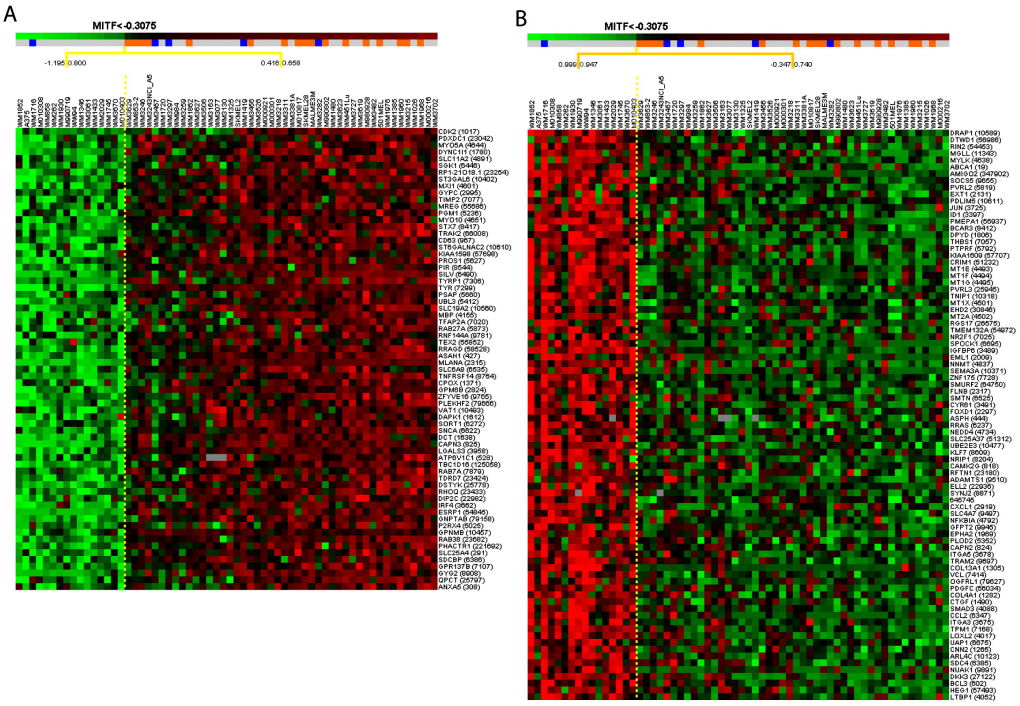


Figure III-7 - MITF associated Modules
A. *MITF* up-regulated module. This module contains genes that were associated with *MITF* and correlate with its expression. The genes are listed with gene symbol and Entrez Gene ID and they are enriched for Vesicular Trafficking and Melanogenesis. **B.** *MITF* down-regulated Module. This module contains genes that were associated with *MITF* and anti-correlate with its expression. The genes are listed with gene symbol and Entrez Gene ID and they are enriched for NFkB/TNF and Invasion/Migration.

MITF expression correlates with targets better than copy number

Expression of *MITF* correlates with the expression of its targets better than *MITF* copy number, though both correlations are statistically significant (p-value of 0.0001 versus 0.04, Figure III-6). This relationship is unidirectional: *MITF* is significantly over-expressed when its DNA is amplified (p-value 0.0004), but over-expressed *MITF* does not always correspond with *MITF* amplification. We find that *MITF* is less correlated with its copy number (rank 294th) than most

other genes in aberrant regions and more than half of the tumors that over-express *MITF* do not have a CNA that spans the *MITF* gene. Comparison of *MITF* target expression between samples with and without *MITF* amplification did not show an effect of DNA amplification on expression of the targets (see Computational methods).

MITF correctly annotated with its known role in melanoma

We used LitVAN to identify the biological processes and pathways represented in each module associated with *MITF*. The module containing the genes most significantly up-regulated by *MITF* (Figure III-7A) is significantly enriched for the terms 'melanosome' and 'pigment granule' (q-value= 10^{-6} for each). It includes targets involved in proliferation such as *CDK2*, consistent with the observation that MITF can promote proliferation via lineage specific regulation of *CDK2*[202]. The module containing genes most strongly inhibited by *MITF* (Figure III-7B) has a metastatic signature strongly associated with invasion, angiogenesis, the extracellular matrix and NF-kappa-B signaling. These modules and their annotation suggest that *MITF* serves as a developmental switch between two types of melanoma, where high *MITF* expression promotes proliferation and low *MITF* expression promotes invasion. Thus our automated, computationally derived findings dissect a complex response and accurately recapitulate the known literature, including the experimental characterization of MITF[193].

The detailed match between the CONEXIC output and empirically derived knowledge of the role of known modulators in melanoma provides confidence in CONEXIC's predictions for modulators that are not well characterized.

Identification of *TBC1D16* as a tumor dependency in melanoma

The second highest scoring modulator identified by CONEXIC is *TBC1D16*, a Rab GTPase-activating protein of unknown biological function. Rabs are small monomeric GTPases, involved in membrane transport and trafficking. TBC1D16 is well conserved and although its targets are not known, a close paralog, TBC1D15, regulates RAB7A (also selected as a modulator, Figure III-4)[203]. We used a module associated with *TBC1D16* to infer its potential role in melanoma (Figure III-8A), and discovered that diverse biological processes are represented by genes in the

module and that more than half are annotated for processes such as melanogenesis, vesicular trafficking and survival/proliferation. This suggests that *TBC1D16* plays a role in cell survival and proliferation.

TBC1D16 is an uncharacterized gene located in an amplified region that contains 23 other genes, including *CBX4*, which is known to play a role in cancer[204]. Expression of *TBC1D16* is not highly correlated with *TBC1D16* copy number, compared to other genes in the region (ranked 7th out of 24) or to all candidate drivers (252th out of 428). Nevertheless, *TBC1D16* is the top scoring gene in the region and the 2nd highest scoring modulator, so it was selected for experimental verification.

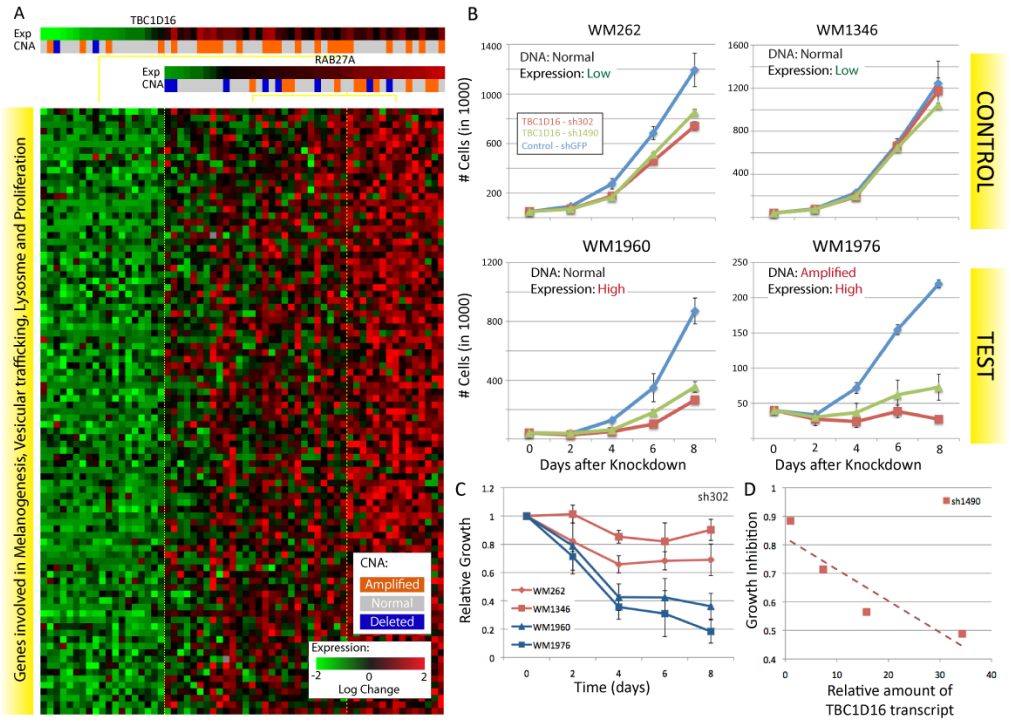


Figure III-8 - *TBC1D16* is necessary for melanoma growth
A. A module associated with *TBC1D16* and *RAB27A*, the genes in the module are involved in melanogenesis, survival/proliferation, lysosome and protein trafficking. **B.** Representative growth curves for each of the 4 STCs infected with *TBC1D16* shRNA, each curve represents 3 technical replicates. RT-PCR was used to confirm that the reduction in the amount of the *TBC1D16* transcript was similar for all of the STCs (Figure III-9). **C.** Change in growth over time, relative to the number of cells plated, averaged over all replicates (see experimental methods). Mean over 3 biological replicates X 3 technical replicates for each STC, see Figure III-9. **D.** Growth inhibition at 8 days is directly proportional to the amount of the *TBC1D16* transcript and is independent of the *TBC1D16* copy number.

The module exhibits a dose-response relationship between *TBC1D16* expression and the expression of genes in the module such that higher expression of *TBC1D16* is correlated with higher expression of genes in the module (correlation coefficient 0.76). These results suggest that knockdown of *TBC1D16* expression in tumors that have high levels of *TBC1D16* will lead to a reduction in proliferation.

TBC1D16 is required for proliferation

To test whether *TBC1D16* is required for proliferation of melanoma cultures we carried out a knockdown experiment. We selected two STCs with high levels of *TBC1D16*, WM1960 (16-fold greater expression than WM1346, DNA not amplified) and WM1976 (34-fold greater expression, amplified DNA) and control STCs, WM262 and WM1346 that express *TBC1D16* at a lower level. We used two shRNAs to knock down *TBC1D16* expression in each of the four STCs and measured growth over 8 days (see experimental methods). RT-PCR was used to confirm that the reduction in the amount of the *TBC1D16* transcript was similar for all of the STCs (Figure III-7). Knockdown of *TBC1D16* expression reduced cell growth in WM1960 and WM1976

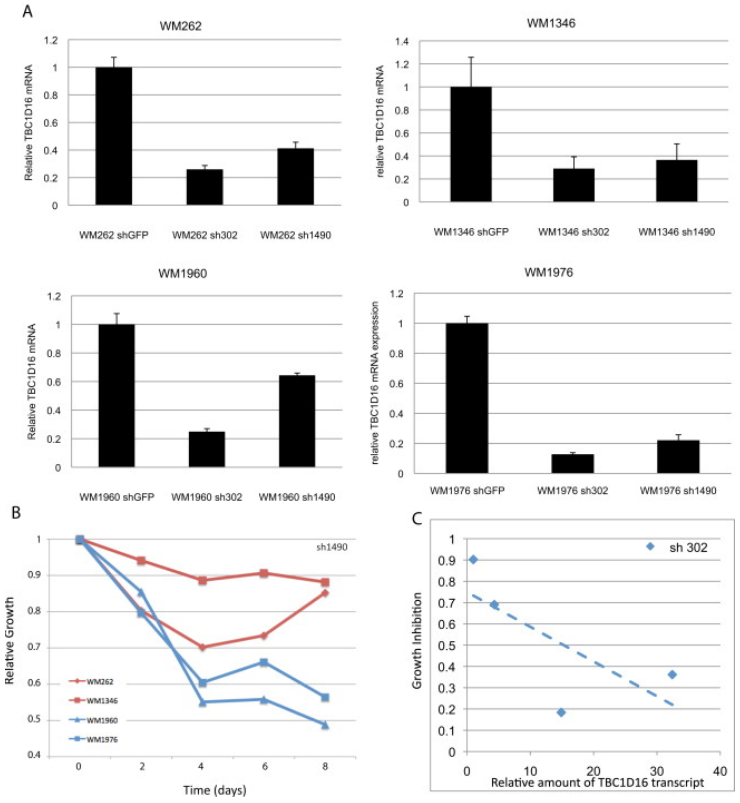


Figure III-9 - TBC1D16 mRNA knockdown and growth effects
A. RT-PCR levels of *TBC1D16* mRNA after shRNA knockdown for each of the 4 STCs tested. The reduction in the amount of the *TBC1D16* transcript was similar in all of the STCs, where each bar represents data from 3 biological replicates. **B.** Representative growth curves for each of the 4 STCs tested with *TBC1D16* knockdown with the hairpin sh1490, each curve represents 3 technical replicates. **C.** Growth inhibition at 8 days is directly proportional to the amount of the *TBC1D16* transcript and is independent of *TBC1D16* copy number. Data averaged on 3 biological replicates X 3 technical replicates for each STC.

to 16% and 40%, respectively, relative to controls infected with GFP shRNA in the same STCs (Figures III-6B, C and D). This result is specific for cultures with high levels of *TBC1D16*, as the controls, WM262 and WM1346, grow at similar rates to cultures infected with shGFP (75%-90%). As predicted, growth inhibition at day 8 is proportional to the amount of the *TBC1D16* transcript and is independent of *TBC1D16* copy number (Figures III-8C and D). Taken together, these results support CONEXIC's prediction that *TBC1D16* is required for proliferation in melanomas that over express the gene.

RAB27A identified and experimentally confirmed as a tumor dependency

The *TBC1D16* module contains a second modulator, *RAB27A*, also known to be involved in vesicular trafficking (Figure III-8A). RAB27A functions, with RAB7A, to control melanosome transport and secretion. RAB7A localizes to early melanosomes, while RAB27A is found in mature melanosomes [205]. CONEXIC selected both *RAB27A* and *RAB7A* as modulators.

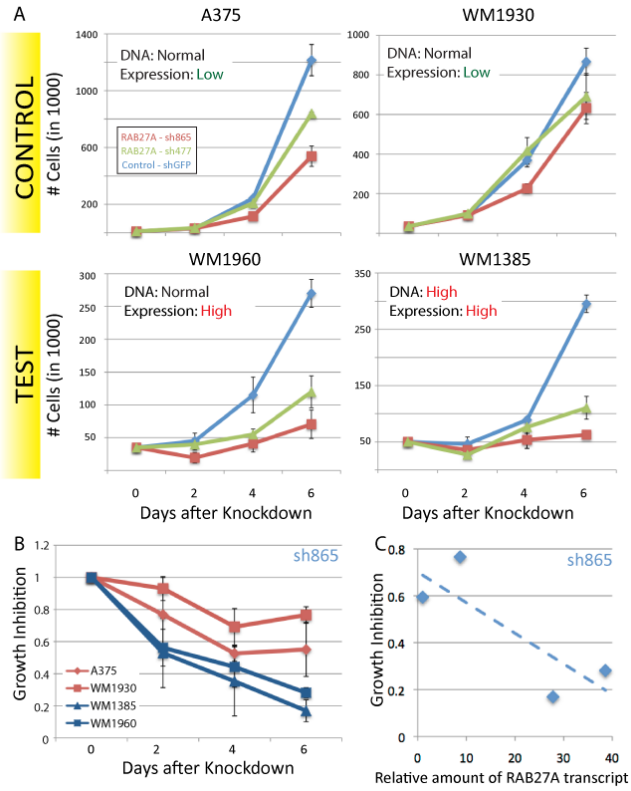


Figure III-10 - *RAB27A* is necessary for melanoma growth
A. Representative growth curves for each of the 4 STCs infected with *RAB27A* shRNA, each curve represents 3 technical replicates. RT-PCR was used to confirm that the reduction in the amount of the *RAB27A* transcript was similar in all of the STCs (Figure III-11). **B.** Change in growth over time, relative to the number of cells plated, averaged over all replicates. Knockdown of *RAB27A* expression in cells that express this gene at high levels reduces proliferation. Data averaged over all replicates for each STC, see Figure III-11 **C.** Growth inhibition at 6 days is dependent on the amount of the *RAB27A* transcript and is independent of *RAB27A* copy number.

RAB27A is in an amplified region that did not pass the standard GISTIC q-value threshold for significance and expression of the gene is not highly correlated with *RAB27A* copy number, compared to other candidate drivers (323th out of 428). Nevertheless, CONEXIC identified it as the top-scoring modulator out of the 33 genes in this region, and ranked it 8th out of 64 modulators and it was therefore selected for empirical assessment.

To test the prediction that *RAB27A* is important for proliferation in tumors with high levels of *RAB27A*, we tested the effect of shRNA knockdown of the *RAB27A* transcript on proliferation. We chose two STCs in which the gene is highly expressed WM1385 (28-fold greater expression compared with A375, DNA amplified) and WM1960 (38-fold greater expression, DNA not amplified) and two controls that express *RAB27A*

at a lower level (A375 and WM1930). Western blots show that expression of RAB27A correlates with expression of the cognate gene in these cultures (data not shown).

Knockdown of *RAB27A* expression using shRNA was similar for all cultures (Figure III-9), but only reduced cell growth significantly in the STCs that overexpress *RAB27A* (18% or 35% in WM1385 or WM1960 relative to the same cultures infected with GFP shRNA). *RAB27A* shRNA had less impact (growth rates of 65-80%) in the control STCs that have low RAB27A (Figures III-10A and B). Growth inhibition at 6 days is correlated with the amount of the *RAB27A* transcript and is independent of *RAB27A* copy number (Figures III-10B and C). Taken together, these results support CONEXIC's prediction that *RAB27A* is a tumor dependency in melanomas that overexpress *RAB27A*.

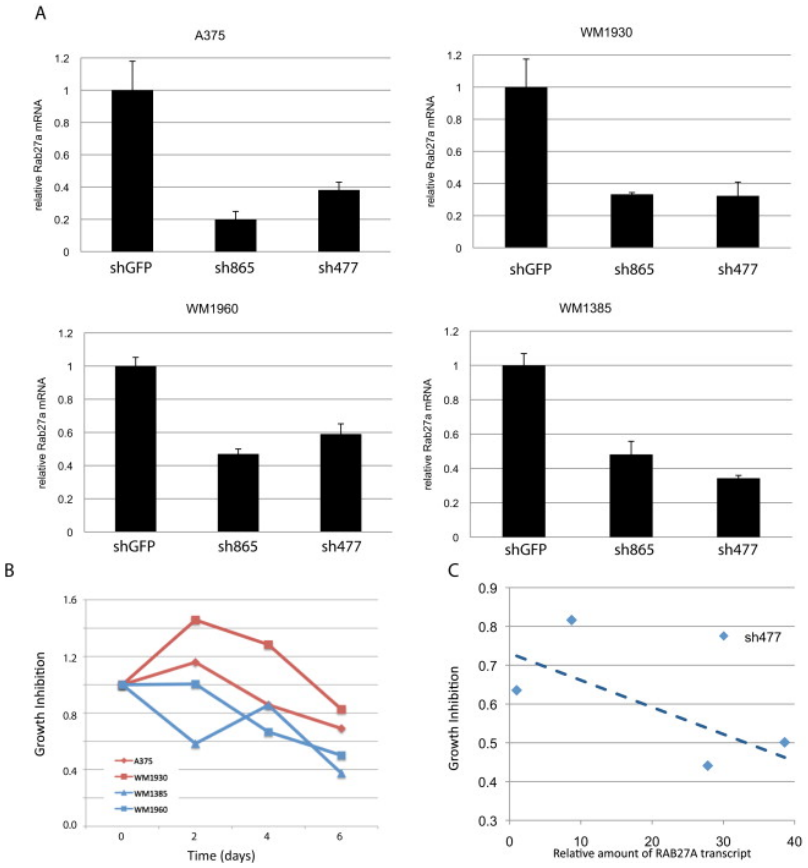


Figure III-11 - *RAB27A* mRNA knockdown and growth effects
A. RT-PCR levels of *RAB27A* mRNA after shRNA knockdown for each of the 4 STCs tested. The reduction in the amount of the *RAB27A* transcript was similar in all of the STCs, where each bar represents data from 3 biological replicates. **B.** Representative growth curves for each of the 4 STCs tested with *RAB27A* knockdown with sh477, each curve represents 3 technical replicates. **C.** Growth inhibition at 6 days is dependent on the amount of the *RAB27A* transcript and is independent of *RAB27A* copy number. Data averaged over all replicates for each STC.

RAB27A affects the expression of genes in associated modules

To test whether *RAB27A* affects the expression of genes in associated modules, as predicted by CONEXIC, we carried out microarray profiling after knockdown of *RAB27A* in the test STCs

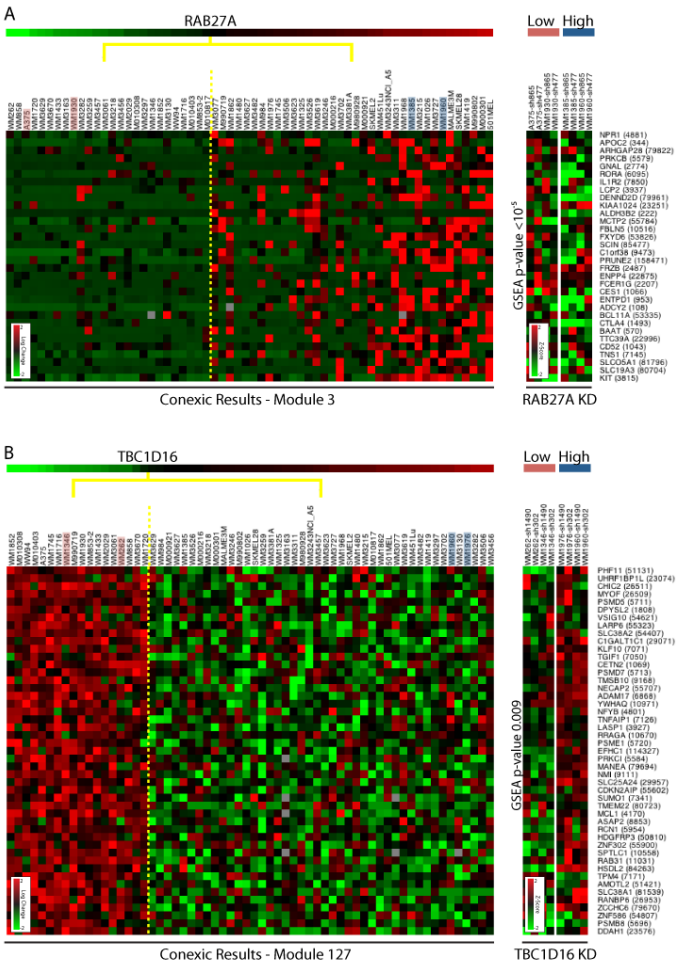
(WM1385, WM1960). We compared the expression profile after *RAB27A* knockdown to a control profile generated by infecting the same STC with GFP shRNA. We used Gene Set Enrichment Analysis (GSEA)[206] to test whether each of the 3 modules associated with *RAB27A* are enriched with genes that are differentially expressed (DEG) after knockdown (see Experimental Methods). We found that all 3 *RAB27A* associated modules are significantly enriched for genes affected by *RAB27A* (p-values < 10^{-5} for all 3 modules, see Figure III-12C), and that these modules responded in the direction predicted by CONEXIC.

These results support our computational prediction that the expression of *RAB27A* affects the expression of the genes in the associated modules. We note that RAB27A functions as vesicular trafficking protein, suggesting that it influences gene expression through an unknown, and likely indirect, mechanism. We used LitVAN to identify the biological processes and pathways represented among the DEGs. Cell cycle related terms are significant among the down-regulated genes, which might be expected given the reduced growth after *RAB27A* knockdown. In addition, we found that genes annotated for the Erk pathway are up-regulated (including *MYC*, *FOSL1* and *DUSP6*). We used GSEA to measure enrichment of an experimentally derived set of genes that respond to MEK inhibition in melanoma[130]. The resulting p-value < 10^{-5} suggests that ERK signaling is altered after *RAB27A* knockdown in these STCs.

TBC1D16 influences the expression of genes in associated modules

We carried out microarray profiling after knockdown of *TBC1D16* to evaluate whether expression of *TBC1D16* affects the expression of genes in the 4 modules associated with it. We used two shRNAs to knock down *TBC1D16* in the test STCs (WM1960, WM1976) and compared the gene expression to controls infected with GFP shRNA (in the same STCs). GSEA analysis established that all 4 modules are significantly enriched for genes affected by differences in *TBC1D16* expression (p-values < 10^{-5}, 0.0002, 0.008 and 0.009 respectively, see Figure III-12). Two modules responded to *TBC1D16* knockdown in the direction predicted by CONEXIC. In addition, GSEA analysis ranked genes in the *TBC1D16* module (Module25) highest out of 177 (based on the GSEA p-value), demonstrating that the genes in this module are the most highly differentially expressed genes in the data set.

The function of TBC1D16 is unknown, but it is predicted to be involved in vesicular trafficking. In our knockdown analysis LitVAN annotated the up-regulated genes with terms related to vesicular trafficking. These include *RAB3C, RAB7A, CHMP1B, RAB18, SNX16, COPB1* and *CAV1*. However, it is not clear how *TBC1D16* affects gene expression or how changes in expression impact vesicular trafficking.



C

RAB27A Moduled Modules

Module	GSEA p-value	Rank
Module 3	$<10^{-5}$	3 ↓
Module 31	$<10^{-5}$	2 ↑
Module 75	$<10^{-5}$	7 ↓

TBC1D16 Moduled Modules

Module	GSEA p-value	Rank
Module 25	$<10^{-5}$	1 ↓
Module 75	0.008	21 ↓
Module 147	$2x10^{-4}$	5 ↓
Module 127	0.009	22 ↑

Figure III-12 - Results of knockdown microarrays for *RAB27A* and *TBC1D16*

A. To the left is one of the modules associated with *RAB27A* and to the right data generated following knockdown (KD) of *RAB27A* for the same genes in the STCs indicated (pink and blue). The expression of genes in the module goes down relative to shGFP as predicted, KD expression heatmap shows Z-scores (see Experimental Methods) showing that these are some of the most differentially expressed genes (DEGs) in the genome. **B.** To the left is one of the modules associated with *TBC1D16* and to the right data

generated following KD of *TBC1D16* in the STCs indicated. The expression of genes in the module goes up relative to shGFP, as predicted. The test STCs (blue) and control STCs (pink) respond differently demonstrating the importance of context (*TBC1D16* over-expression status) in determining the response. **C.** GSEA p-value and ranking (relative to 177 CONEXIC modules) for *RAB27A* and *TBC1D16* associated modules. GSEA was calculated using the median of 4 profiles (2 cell lines X 2 hairpins) on the test STCs. Significant p-values indicate that knockdown of *RAB27A* and *TBC1D16* each affect the subset of genes predicted by CONEXIC (note that 10^{-5} is the smallest p-value possible given that 100,000 permutations are used). The color of the module name represents the predicted direction of response to knockdown (red and green represent up and down regulated, respectively). The arrow represents the observed response to knockdown. The direction of response was correctly predicted for 2/4 *TBC1D16* modules and for all *RAB27A* modules.

Experimental Methods

Data and Processing

We used copy number data for 101 melanoma samples generated by Lin et al.[187]. The SNP locations were translated from HG17 to HG18 using the UCSC liftOver application[207]. Gene expression was available for 62 of these samples using HT-HGU133A Affymetrix chip, which measures the expression of 12725 genes[187]. We removed probe sets whose standard deviation was smaller than 0.25 on a log2 scale, resulting in 12,101 probe sets measuring 8,243 unique genes. We merged probe sets for genes if these agreed and removed inconsistent genes, resulting in a final set of 7,981 genes. Expression values were normalized to mean of zero and a standard deviation of one for each gene.

Experimental Methods

Melanoma short-term cultures (STCs) derived from metastatic foci[187] were cultured in RPMI medium (MediaTech) supplemented with 10% fetal bovine serum (Gemini). A375 melanoma cell line and 293T virus packaging cells were cultured in DMEM medium (MediaTech) with 10% fetal bovine serum. All cells were maintained in 100 units/mL penicillin, 100 µg/ml streptomycin at 37°C under 5% CO_2.

Knockdown was carried out by infection with lentivirus using RNAi sequences designed by the RNAi Consortium. shRNA lentivirus were prepared according to TRC protocols (http://www.broadinstitute.org/rnai/trc), with minor modifications. Cell proliferation assays, RT-PCR, microarrays and immunoblotting were carried out using standard techniques. Primer sequences and detailed methods can be found in Supplementary Experimental Procedures.

All primary data are available at the Gene Expression Omnibus (GSE23884).

shRNA

shRNA knockdown sequences for TBC1D16 and RAB27A were those designed by the RNAi Consortium (TRC):

shRAB27A_865 (TRCN0000005296): CGGATCAGTTAAGTGAAGAAA

shRAB27A_477 (TRCN0000005298): CAGGAGAGGTTTCGTAGCTTA

shTBC1D16_302 (TRCN0000061889): CCTGTGCTTGTACATGGAGAA

shTBC1D16_1490 (TRCN0000061891): GCGAAAGGAGTACTCTGAGAT

The negative control was

shGFP: GCAAGCTGACCCTGAAGTTCA

The shRNA lentiviruses were produced according to TRC protocols (http://www.broadinstitute.org/rnai/trc). Briefly, $2x10^6$ 293T cells were seeded in 100 mm plates, and at 24 hours were co-transfected with 3 µg of pLKO.1-shRNA, 2.7 µg of Δ8.9, and 0.3 µg of VSV-G vectors using Lipofectamine 2000 (Invitrogen) according to the manufacturer's instructions. At 48h post-transfection, viral supernatant was collected, passed through a 0.45 µm filter (Nalgene), and stored in small aliquots in -80°C until use. To reduce nonspecific viral cytotoxicity, each viral supernatant was titrated: Target cells were seeded in 6-well plates ($1x10^5$ - $2x10^5$/ well), and at 24h post-seeding, were infected with dilutions (1:5, 1:10, 1:20, 1:40, 1:100, 1:200) of each lentivirus preparation. Infections were carried out in duplicate, in 2 ml medium/ well containing 6 µg/ml polybrene for 6h. 48h post-infection, puromycin was added to one set of infected cells, leaving the remaining set unselected. Cells were washed with PBS, trypsinized, and collected for counting using a Vi-Cell XR Cell Viability Analyzer (Beckman Coulter) after 3-4 days of selection. For each virus, the ratio of cells surviving in media with puromycin versus media without puromycin was determined for each dilution. Titers that yielded approximately 50% survival were used for subsequent infections.

To perform knockdown experiments, target cells ($1.2x10^6$ - $3x10^6$) were plated the day before infection to obtain 30-40% confluence at the time of infection. Cells were then incubated with virus at the dilution established above, the virus-containing media was removed and fresh media was added. The next day, puromycin was added to select for infected cells (WM1960 and WM1385 at 2 µg/ml; WM1346, WM1976, WM1930, WM262 and A375 at 1 µg/ml). After selection, cells were plated for proliferation assays or used for RT-PCR or immunoblotting.

Cell proliferation assays

Cultures with stable expression of each shRNA construct were seeded in triplicate (technical replicates) in 12-well plates at $1x10^4$ - $5x10^4$ cells/ well in 1ml of medium. Cells were washed with PBS, trypsinized and counted using a Vi-Cell XR Cell Viability Analyzer (Beckman Coulter) or Coulter particle counter (Beckman Coulter) at the times indicated.

Quantitative reverse transcription-PCR analysis

Total RNA was harvested using the RNeasy Mini Kit (Qiagen), and cDNA prepared with the SuperScript III First-Strand Synthesis Supermix Kit (Invitrogen) according to the manufacturers' recommendations. All real-time PCRs were performed in triplicate using PCR SYBR Green Master Mix (Applied Biosystems) on an ABI 7300 (Applied Biosystems). The data were normalized to *TBP*. Gene-specific primer sequences follow:

TBP
forward: 5'- CCACTCACAGACTCTCACAAC-3'
reverse: 5'- CTGCGGTACAATCCCAGAACT-3'
RAB27A
forward: 5'- GAAACTGGATAAGCCAGCTACAG-3'
reverse: 5'- ATATTTCTCTGCGAGTGCTATGG-3'
TBC1D16_*114*
forward: 5'- CTACTCCAAGAACAATGTCTGCG-3'
reverse: 5'- GCCTCTGGATGCGAGAGTTG-3'
***TBC1D16*_1211**
forward: 5'- CGCCCCCGATAAGACATGC-3'
reverse: 5'- CCTTCCGCAGCTTGTACTC-3'
***TBC1D16*_1752**
forward: 5'- GATGAGTCAGACACCTTC-3'
reverse: 5'- GGTACAGCAGTTGTTTCT-3'

Microarray analysis

Total RNA was harvested using the RNeasy Mini Kit (Qiagen). The Affymetrix GeneChip Human Gene 1.0 ST Arrays were used for gene expression profiling, performed by the Dana Farber Cancer Institute facility according to the manufacturer's protocols.

Immunoblot analysis

Western blotting was carried out using standard methods on cell lysates, normalized for total protein. Primary antibodies used were RAB27A (Santa Cruz, 1:500), TBC1D16 (Novus Biologicals, 1:500), β-actin (Sigma-Aldrich, 1:12,000) or α-tubulin (Cell Signaling Technology,

1:1,000). Secondary antibodies were anti-rabbit or anti-mouse IgG, HRP-linked; Cell Signaling Technology, 1:1,000 dilution). The signal was detected using SuperSignal West Pico or West Dura Chemiluminscent Substrate ECL reagent (Thermo Scientific).

LitVAn

LitVAn - Literature Vector Analysis - is an automatic literature-based analysis tool for inference of gene module functionality. The basic principle is similar to other gene set enrichment methods, identifying over-represented terms associated with a subset of genes.

We use the NCBI database, which associates each gene with manually curated papers. Our corpus contains around 70,000 full-text papers. The algorithm is based on TF*IDF score, which gives a higher score to words which are overrepresented in a subset of documents relative to the full corpus. Inverse Document Frequency (IDF), gives each "term" (a word) a score based on the portion of documents it appears in, with high scores for low coverage. Term Frequency (TF), is calculated for a subset of documents rather than the entire set, and is a direct count for the number of times the term appears in the subset. For each set of genes (a module), we count the term frequency in papers associated with these genes and compare this count to the null distribution, using a TF*IDF score[208].

The TF*IDF score takes the "bag of words" approach, ignoring the order of which the words appear and their location in the text (headers, legends, etc.). The documents are first processed by a semantic stemming algorithm, which converts words to their most basic form in order to treat different forms of the same word equally. The IDF score is calculated once for the entire compendium and stored, and the TF score is calculated for each module separately, using all the papers linked to genes in the module. To avoid biasing the module score with terms related to only one gene that has many papers associated with it, we use a "Leave-One-Out" score. With this approach, for each term we indentify the gene that contributes the most to it, and remove its contribution from the TF score of that term. Although TF is generally defined as a linear score, our tests show clear advantage of using a log2 scale.

To evaluate the significance of the TF*IDF score, we generate an expected score based on random modules. Sets of genes in different sizes, varying from 5 to several hundred genes, were randomly selected and their best LitVAn result was used to determine the expected score (calculated using a linear regression between the top TF*IDF score and the number of papers associated with the random module). Based on the randomized scores, the 95% confidence

intervals of the linear regression were used to determine the threshold of significance for a given number of papers.

The output of LitVAn is a ranked list of terms with an indication of their significance level, as well as a map linking the genes, significant terms and papers that contribute to the score. An online version of LitVAn is available at http://litvan.bio.columbia.edu.

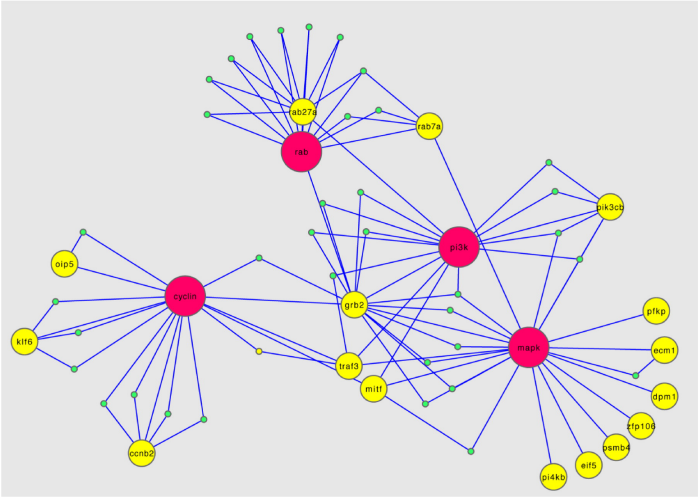


Figure III-13 – A graph output from LitVAn
LitVAn was used to analyze literature trends in the 30 selected modulators appearing in Figure III-4, the top four most significant terms: 'MAPK', 'PI3K', 'cyclin' and 'RAB'. The figure represents the graphic output of LitVAn, where significant terms (red circles) are associated (graph edge) with multiple genes in the query (yellow dots). An edge with a green dot represents a publication that significantly associates between the gene and the term in itself (typically through repeated use of the term throughout the paper) and the interactive version includes a link to the PubMed abstract. Only the top 5 most significant papers for each term-gene combination are represented (if any).

Discussion

We have demonstrated that combining tumor gene expression and copy number data into a single framework increases our ability to identify likely drivers in cancer and the processes affected by them. Gene expression allows us to distinguish between multiple genes in an amplified or deleted region (many of which are indistinguishable based on copy number) and to identify those that are likely to be drivers. The combination of data types allows us to identify regions that would be overlooked using methods based on DNA copy number alone.

Expression of a driver, not its copy number, drives phenotype

The novelty of our method and the key to its success is our modeling paradigm: the expression of a driver should correspond with the expression of genes in an associated module. Examination of MITF and its targets supports our assumptions. Expression of *MITF* best correlates with the expression of its targets, but *MITF* overexpression does not always correspond with *MITF* amplification. A change in DNA copy number is only one of many ways that gene expression can be altered. For example, *MITF* expression can be upregulated via signaling from the Ras/Raf (oncogenic BRAF occurs frequently in melanoma)[209] and Frizzled/Wnt pathways [19].

Most methods for identifying drivers within aberrant regions focus on genes whose expression is well correlated with the copy number of the cognate DNA [187,210]. The expression of many of the predicted drivers we identify is poorly correlated with their copy number, relative to other genes in the region and to all other candidate drivers *MITF* (294th), *TBC1D16* (252th) and *RAB27A* (323th). We believe the discrepancies between CNA and expression arise because there are multiple ways to up or down-regulate a gene. For example, *TBC1D16* and *RAB27A* were both identified as transcriptional targets of MITF [193,211], and are therefore up-regulated when *MITF* is over-expressed. Moreover, we postulate that many drivers are less correlated with their copy number than passengers due to selective pressure; if there is a fitness advantage to up or down regulate expression, the tumor will find a mechanism to do so.

Association between modulator and module

A key feature of our approach is that CONEXIC goes beyond identifying drivers. By associating candidate drivers with gene modules and annotating them using information from the literature, CONEXIC provides insight into the physiological roles of drivers and associated genes. We used LitVAn to find biological processes and pathways overrepresented in each module and to associate drivers with functions, accurately identifying targets of *MITF* and annotating the functions of known drivers (*MITF*, *CCBN2* and *TRAF3*).

The results of microarray profiling following knockdown further support the association between modulator and module and confirm our ability to identify genes affected by *TBC1D16* and *RAB27A*. We successfully connected genes involved in vesicular trafficking to their effects on gene expression, likely through a cascade of indirect influences. In addition to profiling the STCs that highly express each of these genes (test STCs), we also profiled two lower expressing STCs (control STCs), in which the effect of knockdown is less detrimental to growth. For *TBC1D16*, there is substantial overlap in the DEGs in the test STCs (p-value < 10^{-22}), but not in the DEGs between control and test STCs (p-value > 0.76). This reflects the complexity of the transformed state and demonstrates that genetic context has a fundamental impact on the effect of a perturbation.

TBC1D16 and *RAB27A* dependency is context-specific

We tested two drivers predicted by CONEXIC with knockdown experiments, and showed that tumors that express either *TBC1D16* or *RAB27A* at high levels are dependent on the corresponding gene for growth. Our results demonstrate that these dependencies are determined by expression of the gene (in both cases), rather than DNA amplification status, further supporting the assumptions underlying our approach. Thus, we not only identify tumor dependencies, but also the context in which these genes are crucial for proliferation.

Our approach is unbiased with respect to protein function and does not incorporate prior knowledge, thus enabling the identification of dependencies in genes involved with vesicular trafficking.

Of the top 30 drivers selected by CONEXIC, three genes (*TBC1D16*, *RAB27A* and *RAB7A*) are known to be involved in vesicular trafficking [203,205]. All of these genes are amplified (DNA) and highly expressed (RNA) in multiple melanomas. There is increasing evidence that genes controlling trafficking play a role in melanoma. Germline variation in GOLgi PHosphoprotein 3 (*GOLPH3*), a gene involved in vesicular trafficking, is associated with multiple cancers[212]. Our data identifies two novel dependencies that are encoded in somatic CNAs, demonstrates the dependency of melanoma on *TBC1D16* and *RAB27A* expression for proliferation and highlights the potential role of vesicular trafficking in this malignancy.

Beyond Melanoma

The challenge of finding candidate drivers is considerable: tumors are heterogeneous, the data are noisy and highly correlated and there are a large number of possible combinations of drivers and genes in modules. Our approach is successful because it couples simple modeling assumptions with powerful computational search techniques and rigorous statistical evaluation of the results at each step.

The principles underlying CONEXIC can be applied to any tumor cohort containing matched data for copy number aberrations and gene expression. The principle of associating any type of mutation (e.g., epigenetic alterations, coding sequence) with gene expression signatures or other phenotypic outputs that differ among samples will be of increasing importance as sequence and epigenetic data accumulates. Not only does this help to distinguish between driving and passenger mutations, but the genes in the associated module can also provide insight into the role of the driver. This approach can be used to identify the genetic aberrations responsible for tumorigenesis and to find those that relate to any other measurable phenotype, such as the resistance of tumors to drugs.

We anticipate our approach of combining gene expression and genetic lesions will make an important contribution towards a basic mechanistic understanding of cancer and in revealing associations of clinical significance. Cancer is a heterogeneous disease in which we are only just beginning to appreciate the importance of genetic background and the myriad ways in which the

cellular machinery can be redirected towards the transformed state. Methods that begin to dissect this complexity move us another step closer to a world where personalized therapies are routine.

Chapter IV - A system analysis identifies synergy between MEK inhibition and interferonα/β in melanoma

Introduction

Other and I have previously shown that cancer is a heterogeneous disease, both in genotype and phenotype, with each tumor harboring hundreds of mutations. In this work I focus on the transcriptional and phenotypic response to targeted therapy. New drugs target frequent driver mutations, such as HER2 in breast cancer and BRAF-V600 in melanoma[127,213], but studies show dramatic variation in the response to these drugs, both *in vitro* and in the clinic[127,214,215]. The molecular mechanisms that underlie this phenotypic heterogeneity, however, are still not well understood.

The phenotypic variability in response to inhibition of a specific pathway suggests that either the downstream targets of the pathway vary between different tumors, or that alternative oncogenic pathways protect the cells. A better understanding of the interactions and dependencies between pathways, and how these differ between tumors, may explain the variability in response to treatment. The ability to infer dependencies and interactions between pathways is significantly enhanced by measuring the response following perturbation[47,216]. Perturbation, whether inhibition or activation of different components of the pathway, breaks correlated patterns into cause and effect, and can reveal crosstalk between pathways. Post-perturbation data can therefore assist in the identification of cell-line-to-cell-line differences that underlie the phenotypic variance in response pathway inhibition.

Here, I focus on the phenotypic variance following MAPK pathway inhibition in melanoma. Seventy percent of melanoma tumors harbor an oncogenic mutation in the MAPK pathway[217], and drugs targeting the pathway have been recently approved with observed clinical success[127]. However, responses to MAPK pathway inhibitors, both of patients and *in vitro*, vary dramatically[127,214]. We use post-perturbation gene expression data to reveal cell-line-to-cell-line

differences in interactions between the MAPK pathway and other pathways, in an effort to elucidate the origins of phenotypic variance.

To characterize the differences related to the MAPK pathway between tumors that could underlie phenotypic variance, I developed computational tools that analyze pre- and post-perturbation gene expression data. I applied them to gene expression data from a panel of 14 melanoma cell lines treated with PD325901, a MEK inhibitor used to inhibit the MAPK pathway. My results show that although all cell lines harbor an oncogenic activation of the MAPK pathway, a vast majority of MAPK pathway targets are *context-specific* - under the influence of the pathway only in a subset of tumors. Importantly, these differences are not seen in cell lines at steady state, and are only revealed upon perturbation.

My computational methods found that the interferon pathway is either on or off in different cell lines, and identified an interaction between the interferon and the MAPK pathways. This interaction suggested synergy between two unrelated therapies for melanoma – Type-I Interferon (IFNα/β) and MEK inhibitor. I validated these findings experimentally and demonstrated that IFNα/β enhances the cytotoxic effect of MEK inhibition, but does so only in cell lines with low basal activity of the interferon pathway. However, cell lines with high basal activity are resistant to both MEK inhibition and its combination with IFNα/β.

I also found that a deletion of the interferon locus, next to p16, is correlated with basal activation of the interferon pathway and with treatment sensitivity. However, I found that the reason for the resistance in cell lines with high pathway activity is not due to differences in the interferon response, but rather due to a failure in the caspase pathway activation. Taken together, these data suggest that the interferon pathway plays an important role in melanoma cell survival, and can be used both to predict response of MAPK inhibition, and to enhance the efficacy of these inhibitors.

My results demonstrate that inhibition of key oncogenic pathways leads to substantially different transcriptional programs in different cell lines. Such differences should be considered when evaluating the phenotypic and clinical consequences of oncogene-inhibition therapy. Moreover, we show that a better understanding of the interactions and activity state of different

pathways would enable clinicians to tailor new and unexpected drug combinations to individual patients based on their genetic profile, which may lead to better clinical responses.

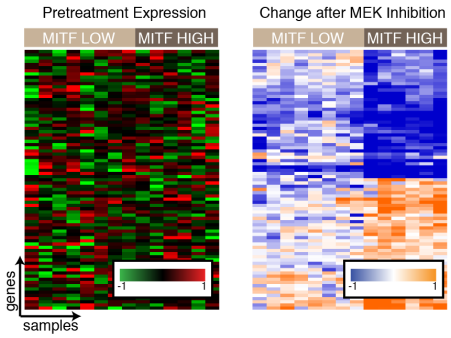
Results

Cell lines and tumors harboring MAPK-activating mutations respond differently to inhibition of the pathway [42]. To characterize the targets and interactions of the MAPK pathway, I chose a panel of 14 genetically diverse melanoma cell lines. This panel represents the spectrum of genetic aberrations in melanoma, including cell lines with NRAS or BRAF mutations, amplifications in *MITF* - a known oncogenic transcription factor in melanoma - and cell lines with AKT activation caused by *PTEN* deletion (figure IV-1A).

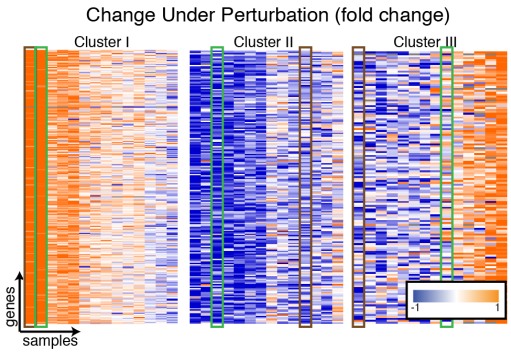
A. Genetic background of the 14 cell line panel

MAPK	PTEN	MITF	Cell lines
BRAF V600	WT	Pos.	MalMe3M, SkMel19 SkMel28, Colo829
		Neg.	A375
BRAF V600	Null	Pos.	A2058, SkMel39
		Neg.	SkMel207, SkMel133 SkMel190, SkMel105
NRAS Q61	WT	Pos.	SkMel30
		Neg.	SkMel2, WM1361

B. Perturbation reveals heterogeneity



C. Context-specific transcriptional targets of MAPK



D. Number of genes with a fold change >= X in Y or more cell lines

Fold change \ Cell lines	1	4	7	10	14
1	3386	936	432	165	18
2	776	137	55	29	4
3	233	36	14	5	2
4	83	12	7	2	1

Figure IV-1 - Heterogeneity in response to MEK inhibition in melanoma
I used a MEK inhibitor, and not the clinically used BRAF inhibitor, to compare the response of BRAF-mut and NRAS-mut cell lines. **A.** BRAF, NRAS, PTEN and MITF status show the genetic diversity of our 14 cell line panel. **B.** MEK inhibition reveals transcriptional targets of MEK (right), undetectable in steady-state conditions (left). These genes are only regulated by MEK in high MITF cell lines. In this and other heat-map figures, columns are samples, and rows are genes. Red-Black-Green represent pre-treatment levels comparing between cell lines, and Orange-White-Blue show expression fold change 8 hours after treatment (both in log2 scale). **C.** 3 gene clusters demonstrating the extent of context-specificity of MAPK targets. The genes in the left and middle clusters are only regulated in a subset of cell lines, while the genes in the right cluster are up- or down-regulated in different cell lines. Cell line order and genes are different in each cluster. As an example, we highlighted A375 (brown) and Colo829 (green) to demonstrate the heterogeneity in response. **D.** Number of differentially expressed genes as a function of fold change and number of cell lines. Arbitrarily choosing the cutoff is poised to mislabel hundreds of genes.

To compare the transcriptional and phenotypic response to MAPK pathway inhibition of both NRAS-mut and BRAF-mut cell lines I used a MEK inhibitor (PD325901, 50nM) that fully inhibits the pathway in all cell lines, and not the more clinically used BRAF inhibitor which works on BRAF-mut cells only (for a comparison of BRAF and MEK inhibitors see materials and methods).

I first characterized the cell lines' cytotoxic and cytostatic responses to MEK inhibition, in each of the 14 cell lines included in our panel. The cell lines display a wide range of phenotypic responses to MEK inhibition, including full or partial cell cycle arrest, and a range of cytotoxic responses, from complete resistance to high sensitivity (figure IV-2). Notably, and contrary to previously published results [42,45], I found that key genetic aberrations common in melanoma, including *MITF* and *PTEN* status, and MAPK mutation type, fail to explain and predict the cytotoxic and cytostatic responses.

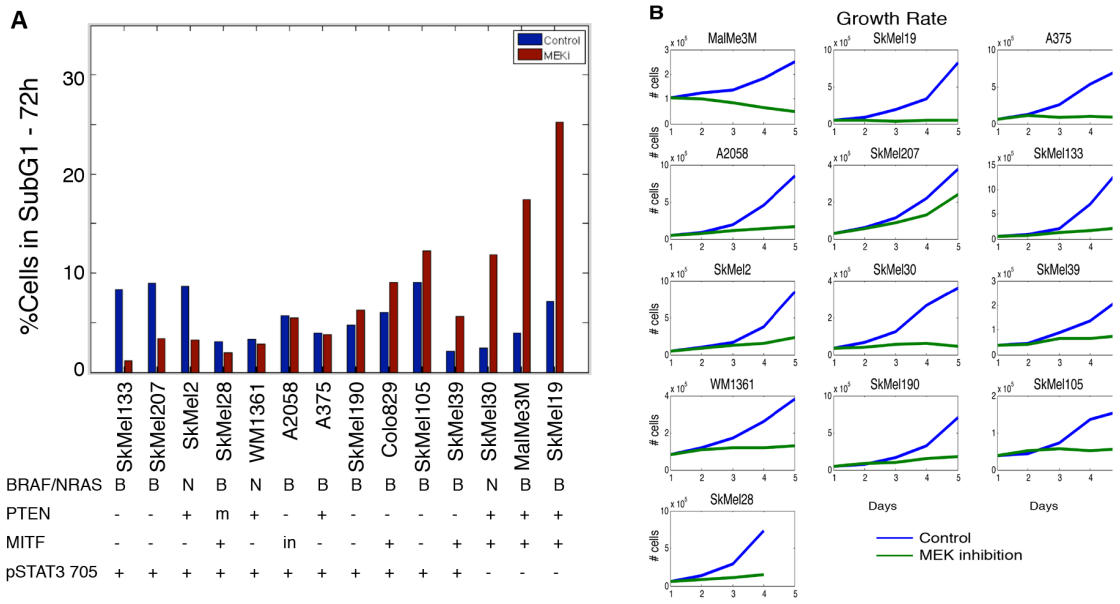


Figure IV-2 - Phenotypic Heterogeneity

B. Growth curves of untreated (blue) and MEK-inhibited (green) cells showing dramatically different responses. Cells were plated in 6-well plates, with 50K cells per well with 2mL of growth media. Treated with DMSO (vehicle) or PD901 after 24h and counted every 24 hours. **C.** Cytotoxic response to MEK inhibition, assessed by percentage of cells in SubG1 72 hours after treatment with DMSO (blue) or PD901 (red). Genotype, MITF and pSTAT3 status listed at the bottom. The 3 cell lines with low STAT3 activation show the strongest cytotoxic response (right-most bars).

Heterogeneity in transcriptional response to MAPK inhibition

I hypothesized that differences in the interactions and downstream targets of the MAPK pathway underlie the observed phenotypic variability in response to MAPK inhibition. To characterize MAPK targets and crosstalk with additional pathways, I measured both pre- and post- MEK inhibition gene expression profiles. I used gene expression 8 hours following MEK inhibition to capture the peak of MEK inhibition before a known feedback loop reactivates the pathway [130].

My data demonstrate that perturbations reveal heterogeneity between tumors that is not observable in steady state expression. For example, a set of genes that display no correlation with respect to mRNA expression levels before pathway inhibition becomes strongly correlated following MEK inhibition (figure IV-1B). Moreover, the data show that some genes, including *CDK2* and *DUSP23*, are either up- or down- regulated by MEK inhibition in different cell lines (figure IV-1C). In fact, most genes are regulated by MAPK in only a subset of the cell lines we tested. For example, only 18 genes change by >2 fold in all 14 cell lines, but 936 genes pass this threshold in 4 or more cell lines (figure IV-1D). I term those genes *context-specific targets*, as they are under the control of MAPK in only a subset of cell lines. These context-specific targets include, among many others, *DUSP1* and *DUSP2*, members of phosphatases that regulate MAPK pathway activity. These genes may play an important role in feedback, but have been overlooked because their expression level changes only in a subset of cell lines (figure IV-3).

Notably, no two cell lines present the same transcriptional response to MEK inhibition, demonstrated by the large variability in gene expression changes (Figure IV-1C). These results are consistent with the idea that the MAPK pathway controls different genes in different contexts and suggest that the interactions of the MAPK pathway vary between cell lines. I postulated that the transcriptional response to MEK inhibition can be used to infer the interactions between MAPK and other

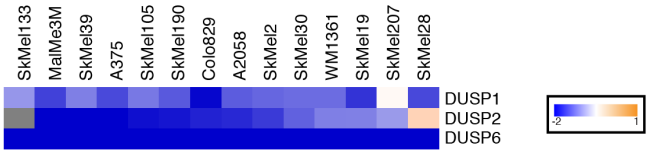


Figure IV-3 - DUSPs are context specific targets
DUSP1 and DUSP2 are context-specific targets of the MAPK pathway in melanoma, while DUSP6 is a target of MAPK in all cell lines.

pathways, and how these differ between cell lines. Therefore, I developed a method to identify targets in a subset of tumors, and not only targets shared by all tumors.

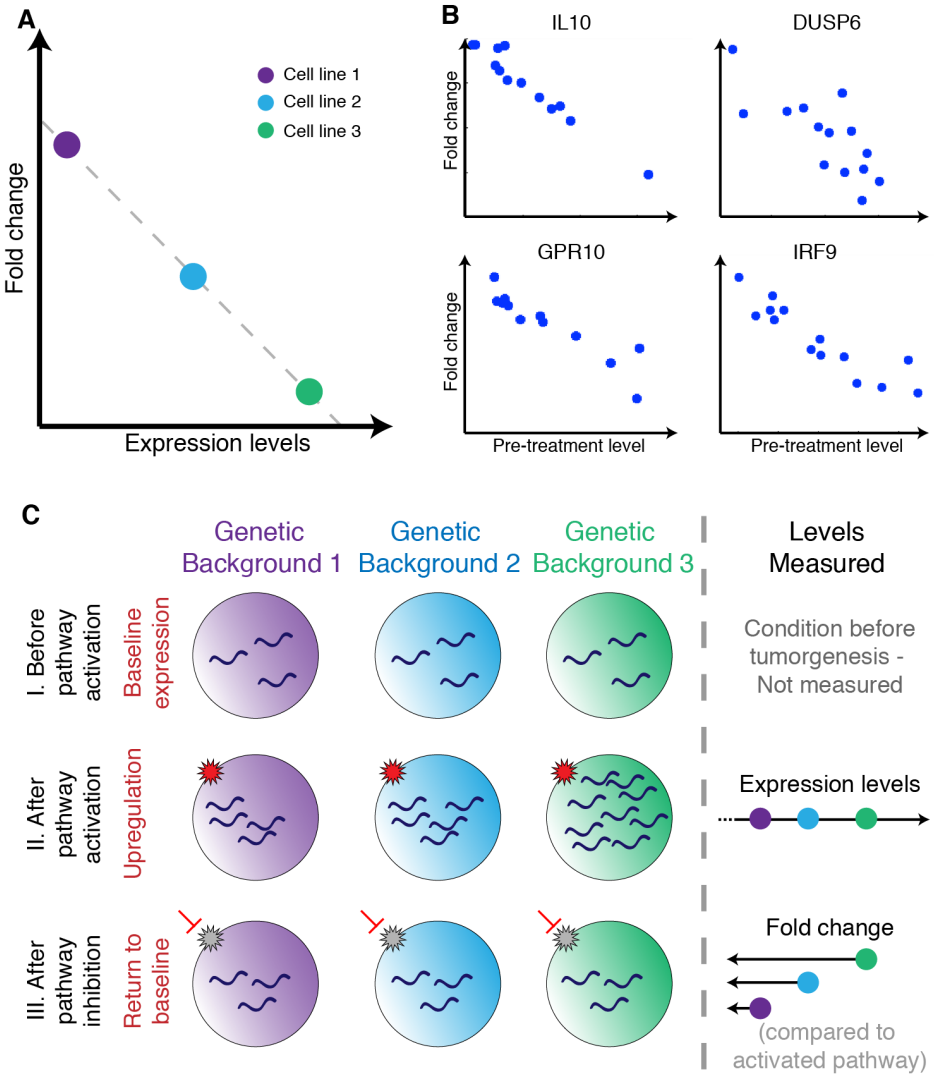


Figure IV-4 - Tippi – a method to identify context-specific transcriptional targets.
A. Tippi – Target Identification using Pre- and Post-Inhibition data. Changes in expression following MEK inhibition are proportional to the pre-treatment expression levels. **B.** Four targets identified by Tippi, including the known target DUSP6. Each dot represents the expression level of a gene in one cell line, x-axis shows the pre-treatment levels, and y-axis is the fold change after MEK inhibition. Tippi identified 793 genes that exhibit such behavior (see figure IV-5). **C.** Tippi's underlying model – I. mRNA expression level of target gene is similar in all tumors' precursor cells. II. Oncogenic pathway is activated to different degree in different tumors, leading to different target gene expression levels. III. Pathway inhibition brings gene expression to its original pre-oncogenic level. Tippi uses the correlation between pre- and post- inhibition expression levels to identify downstream targets, and is independent of the extent of fold change and the number of cell lines in which the gene is under the control of MAPK.

Target Identification using Pre- and Post-Inhibition data

I developed a method – Tippi (Target Identification using Pre- and Post-Inhibition data) - to identify context-specific transcriptional targets. Tippi identifies targets with strong anti-correlation between their steady state expression and fold change levels across cell lines (Figure IV-4A). Several known downstream genes of MAPK, such as *DUSP6* and *IL10*, exhibit such pattern (figure IV-4B). Overall, Tippi identifies 793 genes as MAPK targets using two independent biological replicates, of which only 5 were previously known.

Tippi is not a threshold-based method. The approach is based on the assumption that the degree of MAPK pathway activation varies between cell lines, leading to different basal gene expression levels in different cell lines. However, after pathway inhibition, the expression of a given gene returns to its pre-activation level, and the extent of change is proportional to the degree of the activation (Figure IV-4C). Therefore, Tippi is independent of fold change thresholds and has no restriction on number of cell lines in which a gene's expression is altered.

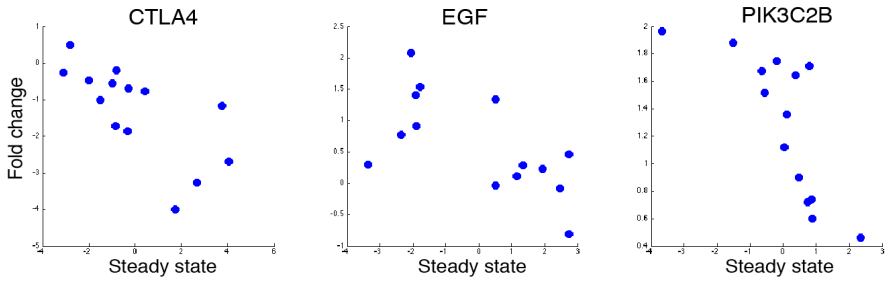


Figure IV-5 – Context-Specific targets
CTLA4, EGF and PIK3C2B, among many other genes, are context-specific targets identified by Tippi.

Tippi's advantages allow it to identify hundreds of context-specific targets of MAPK that are missed by threshold-based methods. For example, *EGF*, *CTLA4* and *PIK3CB*, that were shown to play important roles in melanoma [218,219], have a relatively low expression change in a subset of cell lines. Hence, they are missed by typical threshold-based methods, but identified by Tippi (figure IV-5). Out of the 793 genes identified by Tippi, only 5 show large changes in their expression level across all 14 cell-lines and were therefore previously identified as MAPK target by threshold-based methods [130]. The other 788 targets either change only in a subset of cell lines, or change to a lesser extent (full algorithmic details under computational methods).

Importantly, Tippi only identifies genes consistent with its assumptions – similar basal expression levels before oncogenic activation and short mRNA half-life of the target. Hence Tippi is not comprehensive; rather, it supplements other methods to detect transcriptional targets.

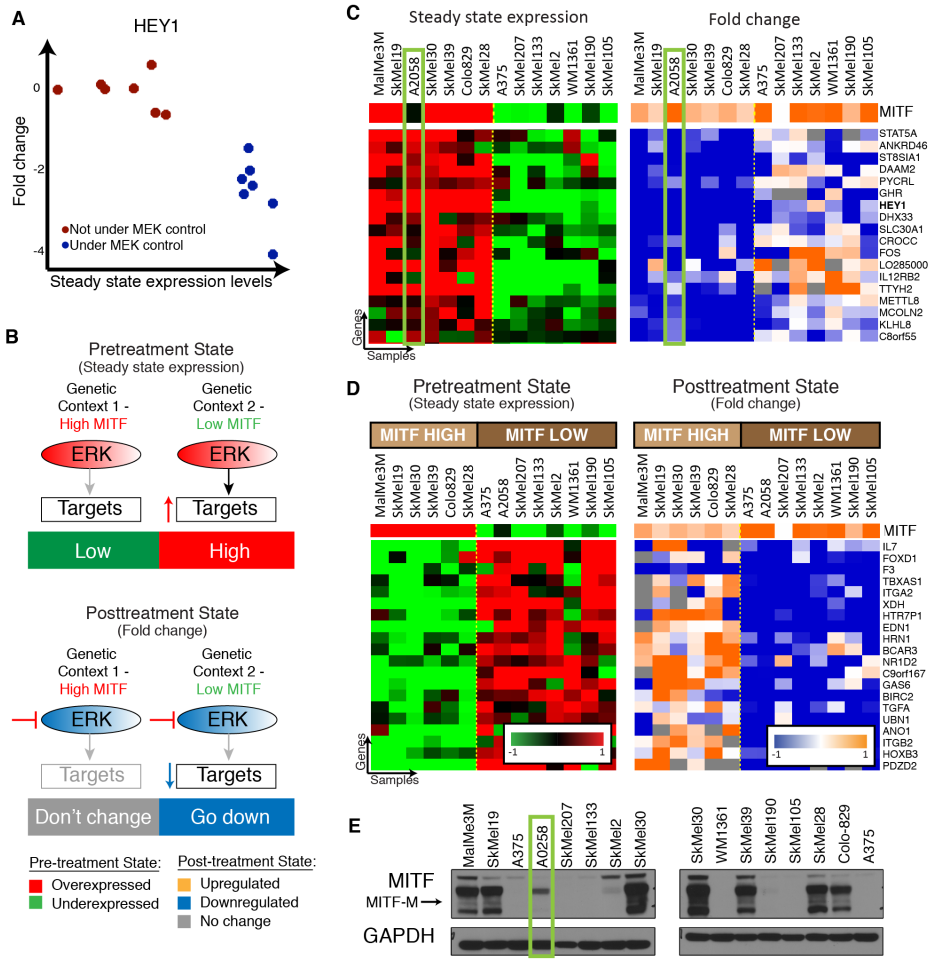


Figure IV-6 - COSPER identifies COntext-SPEcific Regulation

A. HEY1, a target identified by Tippi, exhibits a context-specific behavior - regulated by MAPK in only a subset of cell lines (blue dots). MEK inhibition doesn't affect its expression in the other group of cells (red), and its basal expression is lower in these cell lines. **B.** A cartoon of context-specific regulation exhibited by HEY1. ERK up-regulates a set of targets in high-MITF cell lines only, while it has no effect in low-MITF lines (upper panel). Therefore, the genes are only affected by MEK inhibition of ERK in high-MITF lines (lower panel). **C.** COSPER identifies gene clusters with context-specific regulation. This cluster contains HEY1, and its genes are controlled by MAPK only in cell lines with MITF-M expression. Heat-map on the left shows expression levels before pathway inhibition, and heat-map on the right shows fold change after MEK inhibition. MITF expression, which is not part of this cluster, is in the top row. Part of the cluster is shown (full cluster in IV-8A). **D.** A second cluster identified by COSPER. Its genes are overexpressed in low-MITF cell lines, and are down-regulated only in these cells after MEK inhibition. MITF expression is in the top row. Part of the cluster is shown (full cluster in IV-8B). **E.** MITF protein levels in all 14 cell lines. A2058 (green rectangle) is the only low mRNA-MITF cell line that expresses the MITF-M isoform.

Tippi identifies prevalent context-specific regulation of MAPK targets

Roughly half of the targets captured by Tippi are under the control of MAPK in only a subset of cell lines. For example, *HEY1*, a transcriptional repressor in the notch pathway (figure IV-6A), has high expression levels before inhibition in one group of cell lines, which decrease after inhibition. In the second group, the basal levels of *HEY1* are lower and do not change after MEK inhibition. Therefore, *HEY1* is under MAPK control in a subset of samples, i.e. its regulation is *context-specific*.

Further evaluation of *HEY1*'s regulation shows that it is under MAPK control only in cell lines with relatively high basal expression levels of *MITF*. MITF is a lineage-specific oncogenic transcription factor [12,103,220], and is regulated by MAPK, as its expression levels increase after MEK inhibition (figure IV-7A,B). Other genes, including a number of known MITF targets [193], also exhibit such behavior (figure IV-6B); they are highly expressed in MITF-high cell lines, and repressed in these lines after MEK inhibition (figure IV-6C).

The patterns of regulation are not restricted to *HEY1*'s pattern, depicted in figure 3B. For example, another group of genes, some of which are known MITF targets [193], exhibit a different behavior - they are highly expressed in high MITF cell lines, but they are upregulated in these same cell lines after MEK inhibition (figure IV-7C). Among context-specific MAPK targets, both the context (grouping of the cell lines that are influenced by MAPK), as well as the pattern of this regulation varies.

COSPER identifies a large number of MAPK targets in melanoma

I reasoned that context-specific targets provide increased resolution about the differences in MAPK pathway dependencies and interactions between cell lines. To identify these context-specific targets, I developed a computational framework – COSPER (Context-SPEcific Regulation) – that detects clusters of genes that respond to MEK inhibition in similar, coordinated fashion, in subsets of samples. In each cluster, the cell lines are divided into two groups, or *contexts*, and the genes behave differently in each context, both before and after pathway inhibition (figure IV-6B, full algorithmic details under experimental methods).

Overall, COSPER identified 70 context-specific clusters with 5 genes or more, and assigned 1024 genes to clusters (genes are allowed to belong to more than one cluster), of which 244 genes were also identified by Tippi. Notably, none of the clusters correlate with the oncogenic activation of MAPK (BRAF or NRAS), or with the cells' *PTEN* status. Moreover, I also explicitly tested for genes correlated with these aberrations, but no genes were found to be significantly associated with these mutations (see computational methods). Instead, I believe that these clusters represent context-specific crosstalk and regulation involving additional pathways and processes.

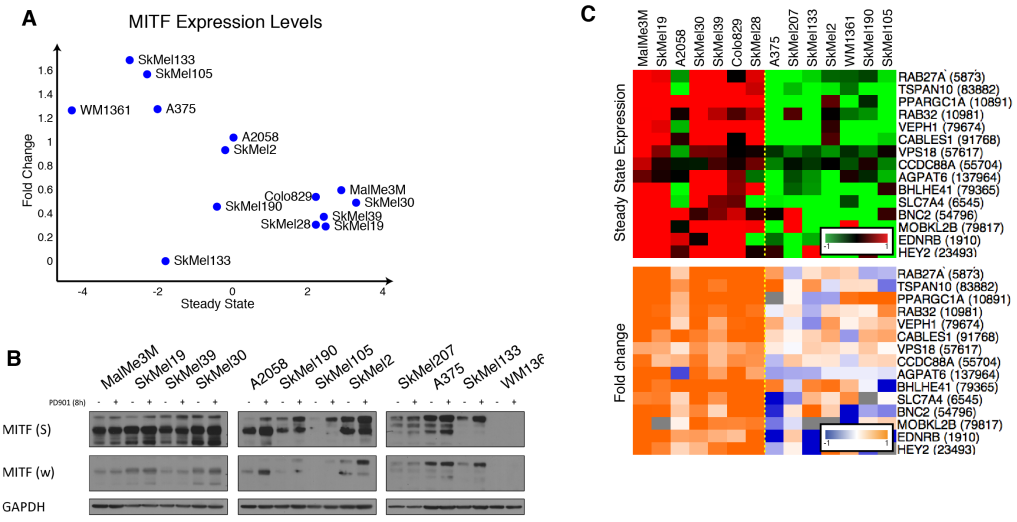


Figure IV-7 – MITF is regulated by MAPK in a context-specific way
A. MITF mRNA expression levels before (x-axis) and after (y-axis) MEK inhibition. Steady state and fold change levels are negatively correlated. **B.** Levels of MITF protein isoforms in 12 cell lines, before and 8 hours after MEK inhibition. Each isoform is regulated to different degrees in the different cell lines. Strong (S) and Weak (W) film exposures are shown. **C.** Another pattern of context-specific regulation identified by COSPER (see figure IV-6C for comparison). The genes in this cluster show high basal expression in high MITF cell lines, and are up-regulated only in these cell lines after MEK inhibition.

Different MITF isoforms give rise to distinct regulatory programs

Fifteen clusters discovered by COSPER, containing 401 genes in total, associate with MITF. The clusters either have a perfect correlation with MITF expression, such as the cluster in figure IV-6D (MITF-expression cluster), or have 1-2 cell lines "switch sides" - they behave similarly to cell lines with the opposite MITF status (figure IV-6C and IV-8A).

Steady state expression levels alone are not sufficient for detection of context-specific targets. For example, 2244 genes are correlated with MITF before treatment, but only ~20% exhibit MITF-dependent behavior following MEK inhibition. Only by combining both pre- and post-inhibition data, was I able to focus our search and identify context-specific targets of MAPK.

I focused on two clusters identified by COSPER. We postulated that clusters with imperfect correlation to MITF levels represent different isoforms of MITF. One cluster is associated with MITF mRNA expression (figure IV-6D and IV-8B), while the other is associated with the abundance of the MITF-M protein isoform (figure IV-6C). A2058 is the only cell line with low mRNA expression that expresses the MITF-M isoform. In melanoma, MITF has at least 4 different protein isoforms [221], as shown by a Western blot (figure IV-6E).

The different functional activities of the genes in the two clusters show that different MITF isoforms regulate different processes. The promoters for genes in the MITF-M cluster are highly enriched for the MITF binding site (CACATG) [119] (pvalue=10^{-3} compared with 0.7 for genes in MITF-expression, see computational methods). However, the MITF-expression cluster, but not the MITF-M cluster, is enriched for the GO annotation “melanocyte differentiation” (qvalue=10^{-4}), suggesting that another isoform of MITF is responsible for cellular differentiation.

These clusters suggest that different combinations of MITF isoforms expressed in each tumor define various MITF states and distinct regulatory programs. Each of these isoforms presumably regulates a different set of genes and processes that are being influenced by MAPK to different degrees (figure IV-7B). These combinations and interactions create a vastly more complicated regulatory network than has been previously appreciated[103], and still remains to be fully elucidated.

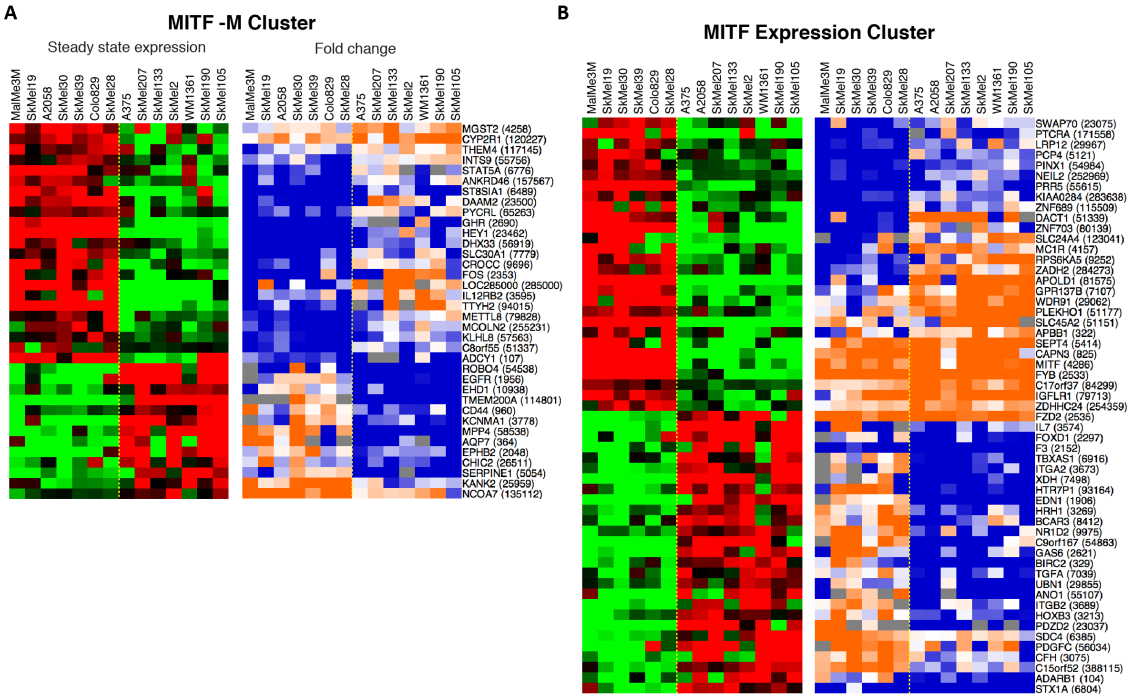


Figure IV-8 – MITF clusters
A, B. Full clusters of the clusters shown in figure IV-6C,D, respectively. Each cluster contains several regulation patterns.

Growth under MEK inhibition correlates with STAT3 and non-canonical NF-kB

The phenotypic response to MEK inhibition varies significantly between cell lines. While MITF levels were previously found to correlate with response to MEK inhibition in melanoma [45], data show that, albeit correlated, not all six MITF-high cell lines are sensitive to inhibition. Specifically, only 3 out of 6 MITF-high cell lines show strong growth arrest and cytotoxic responses following treatment (figure IV-2A,B). I used COSPER to find pathways that can better predict response to MAPK inhibition.

COSPER identified one cluster that separates the 3 cell lines with the strongest growth inhibition from the other 11 cell lines (figure IV-9A and IV-10A). This cluster's genes can be used to identify potential upstream pathways that control their expression. Gene set enrichment analysis associates the genes with cytokine-cytokine receptor pathway (qvalue<10^{-3}), and with miR-19 and miR-17 (qvalue<10^{-3}). Both gene sets are associated with STAT3 [222,223]. Additionally, the cluster also includes several genes associated with the NF-kB pathway (qvalue<10^{-3}).

These predictions were confirmed by measuring STAT3 and NF-κB activity in the cell lines. Levels of pSTAT3-Y705, an indicator for STAT3 activity, and nuclear localization of two non-canonical NF-κB proteins, RelB and p50, but not the canonical NF-κB proteins, match the cluster's contexts (figure IV-9B-C, IV-10B). Sensitive cell lines present low activity of both pathways. It has been previously reported that activation of p50 predicts worse clinical outcome [224]. Taken together, it is possible that activation of these pathways gives cells partial protection from the cytotoxic and cytostatic effects of MEK inhibition, and targeting those pathways can be beneficial for patients.

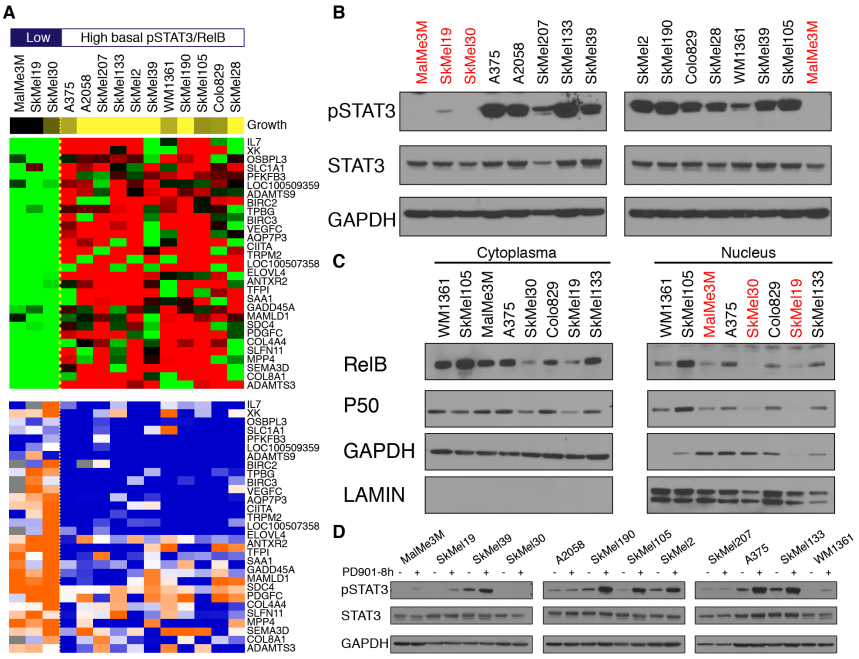


Figure IV-9 - STAT3 and non-canonical NF-κB basal activity levels predict growth phenotype

A. COSPER cluster correlated with growth rate under MEK inhibition. Growth rate shown in the top row, yellow represents the ratio between the number of cells after 4 days of treatment with number of cells before treatment. 3 Cell lines (left), with low gene expression levels in cluster genes, are more sensitive to MEK inhibition. The cluster is enriched for STAT3- and NF-κB- related annotations. **B.** pSTAT3-Y705 levels are correlated with the cluster, as predicted by COSPER. The 3 highly sensitive cell lines (in red) have low pSTAT3. **C.** Sensitive cell lines (in red) have lower RelB and P50 activity, but not of other NF-κB proteins (figure IV-10B). NF-κB activity was assessed by nuclear localization (right panels). **D.** STAT3 is directly regulated by MAPK. MEK inhibition upregulates pSTAT3-Y705 levels.

Western blots confirmed COSPER's prediction regarding network state and showed that NF-κB and STAT3 are either on or off in different cell-lines. However, COSPER also predicts a direct interaction between MAPK and those pathways, as the post-inhibition expression data indicate that expression of NF-κB and STAT3 targets are altered following MEK inhibition. I therefore tested whether STAT3 and NF-κB themselves are also regulated by MAPK, and found that STAT3 activity, but not NF-κB activity, is regulated by MAPK, as pSTAT3 levels increase following MEK inhibition (figures IV-9D and IV-10C).

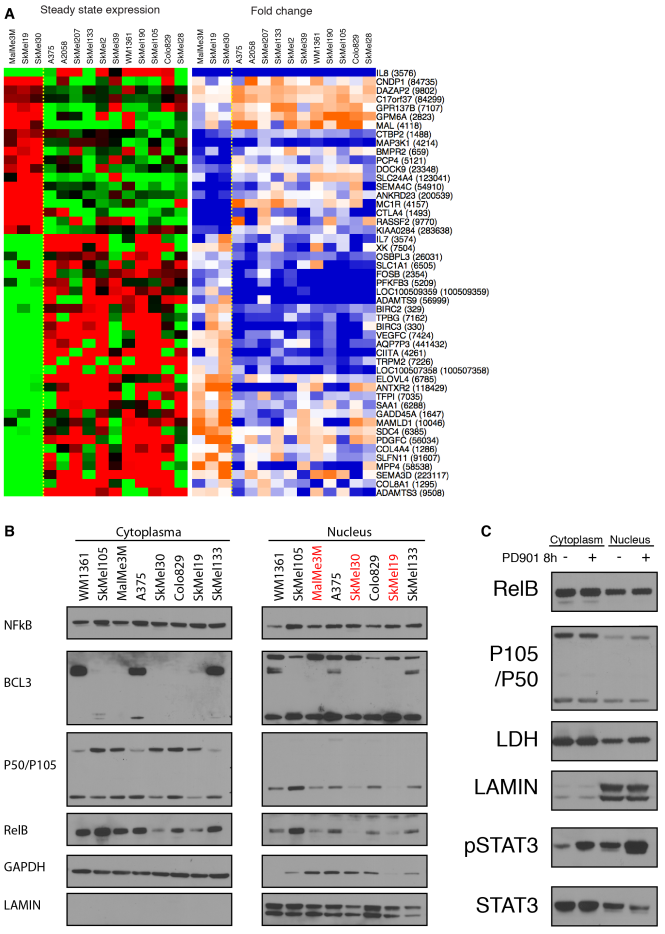


Figure IV-10 - NF-kB regulation
A. Full cluster shown in figure IV-9A. The cluster is enriched for STAT3 and NF-kB targets, and displays several regulation patterns. **B**. Canonical NF-kB genes – NFkB and BCL3, are not correlated with growth phenotype (sensitive lines shown in red). **C.** MEK inhibition does not lead to an upregulation in P50 and RelB activity (nuclear localization) and expression levels. Increase in pSTAT3 levels shown for comparison.

Using the MITF and STAT3/NF-κB examples, I showed that COSPER infers both network state and interactions between pathways. However, when running COSPER on steady-state data alone, the resulting clusters are much larger, less specific, and therefore less informative than the clusters resulting from using both datasets (see computational methods). Therefore, post-inhibition mRNA expression data plays a critical role in identifying the state and interactions of the pathways, and how they are associated with phenotypic response. Although only 14 cell lines were used,

COSPER was able to identify these two pathways and rule out associations, such as MITF levels, made using much larger control-only panels [45].

Both low- and high-pSTAT1 cell lines are resistant to interferon

After successfully using COSPER to identify the growth inhibition phenotype, I looked for additional clusters that might predict response to treatment. I identified a small cluster correlated with basal activity of the Type I interferon pathway. Since interferon (IFN) is one of the few approved drugs for metastatic melanoma, I decided to focus on this cluster.

This cluster contains several known interferon stimulated genes, *IRF7*, *IRF9*, *CCL5* and *IFI44L* (figure IV-11A), and splits the cell lines into two groups; the first contains 3 cell lines with an up-regulation of interferon response genes, while cell lines in the second context express these genes at lower levels. Notably, the cell lines with up-regulation of the STAT1-interferon response genes are not the same 3 cell lines with low activity of STAT3 and NF-κB. Levels of pSTAT1-Y701, an indicator of the interferon-STAT1 activity levels [225], confirmed that the high basal expression levels of the pathway targets correspond with high signaling activity of the pathway (figure IV-11B).

High basal activity of the STAT1-interferon pathway has been previously shown to be necessary, but not sufficient, for IFNa/b-induced apoptosis [226]. To test this claim, 3 low- and 3 high-pSTAT1 cell lines were treated IFNβ and apoptosis levels were assessed by TUNEL. All low-pSTAT1 and 2 high-pSTAT1 cell lines were resistant to the cytotoxic effects of IFNb, and one high-pSTAT1 cell line was marginally sensitive (figure IV-11D). Both IFNα and IFNβ were tested, and as previously shown [227], IFNβ led to a greater apoptotic response than IFNα (figure IV-12A); thus, IFNβ was chosen for further analysis. My results confirmed the previous findings that STAT1 activity is necessary, but not sufficient, for IFNα/β sensitivity.

MEK inhibition and IFNβ treatment act synergistically to increase apoptosis

The expression data indicate that MEK inhibition leads to an up-regulation of the IFNα/β pathway. Analysis of protein levels by Western blot demonstrates an increase in pSTAT1 levels after MEK inhibition, confirming a crosstalk between MAPK and STAT1 (figure IV-11c). Because interferon activity was shown to be required for IFN-induced death, I hypothesized that IFN might synergize with MEK inhibition to increase apoptosis.

First, the cytotoxic effect of MEK inhibition on both high- and low-pSTAT1 cell lines was assessed. I found that high-pSTAT1 cell lines are mostly resistant to inhibition of the pathway,

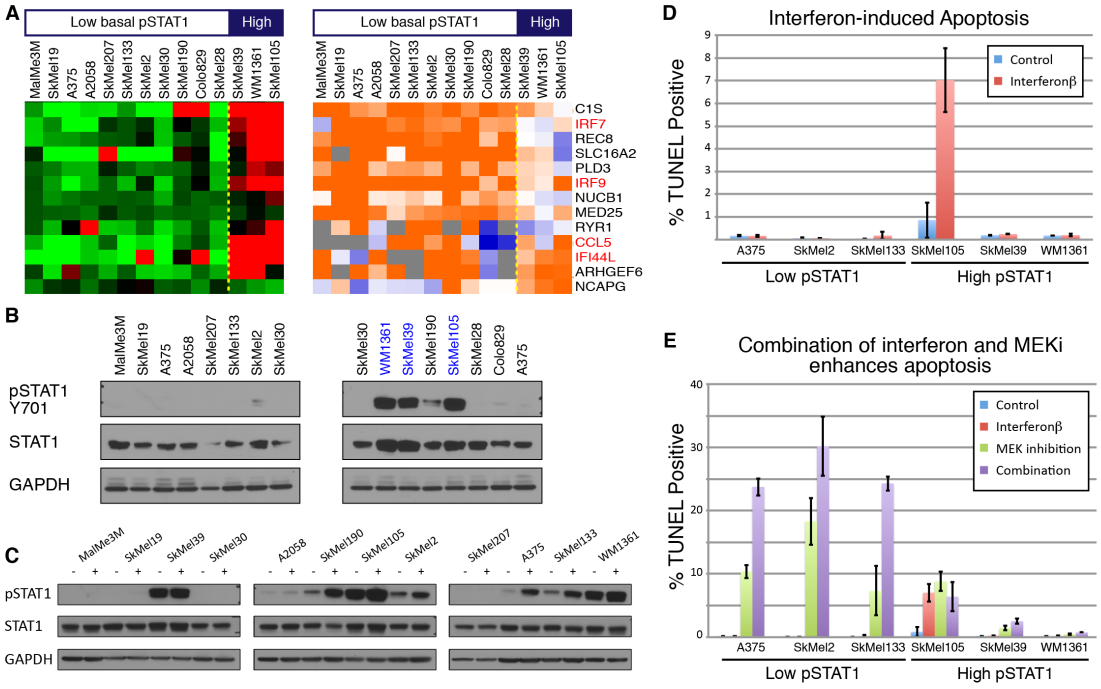


Figure IV-11 - IFNβ enhances cytotoxic response of MEK inhibition in low-pSTAT1 cell lines
A. COSPER identified a cluster containing several known interferon targets marked in red. Three cell lines have high expression, and MEK inhibition upregulates the pathway in the other 11 cell lines. **B.** pSTAT1-Y701, a marker for interferon-STAT1 pathway activity, is correlated with the gene expression and shows high basal activation level in the 3 high cell lines (blue). **C.** MEK inhibition leads to up-regulation of pSTAT1 in all cell lines. **D.** High interferon pathway activity is necessary, but not sufficient, for IFN-induced death. Only one out of 3 high-pSTAT1 cell lines respond to IFNβ (red), using TUNEL staining as a marker for apoptosis 72 hours after IFNβ treatment. I used IFNβ, and not IFNα, due to its higher efficacy (see figure IV-12A). **E.** MEK inhibition induces death in low-pSTAT1 cell lines only (green). IFNβ and MEK inhibition in low pSTAT1 cell lines synergize to increase apoptosis levels (purple). High pSTAT1 cell lines show only mild response to the MEK inhibitor and its combination with IFBβ (right). IFNβ alone and untreated cells (red and blue respectively) have almost no cytotoxic response.

while low-pSTAT1 cells are sensitive (figure IV-11e). Notably, both groups contain NRAS and BRAF mutant cell lines, and cell lines with high and low MITF expression, although both MITF-low cell lines and NRAS mutant cell lines have been previously reported to be less sensitive to MAPK pathway inhibition [45,126]. Moreover, the results also show that the cytotoxic response of MEK inhibition is independent of its cytostatic response. For example, SkMel133, one of the only cell lines that continue to grow rapidly under MEK inhibition (figure IV-12B), has relatively high apoptosis levels under MEK inhibition.

I then examined the cytotoxic effect of the combination of MEK inhibition and IFNβ. While IFNβ as a single agent has no cytotoxic effect on low-pSTAT1 cell lines, it notably enhances the cytotoxic response of MEK inhibition, increasing TUNEL-positive cells by almost two-fold (figure IV-11E, IV-12B). While low-pSTAT1 cell lines show a strong sensitivity to the combination of MEK inhibition and IFNb treatment, high-pSTAT1 cell-lines have consistently low apoptosis levels (figure IV-11E). Moreover, these cell lines seem to be resistant to the cytotoxic effects of both MEK inhibition alone and the dual treatment.

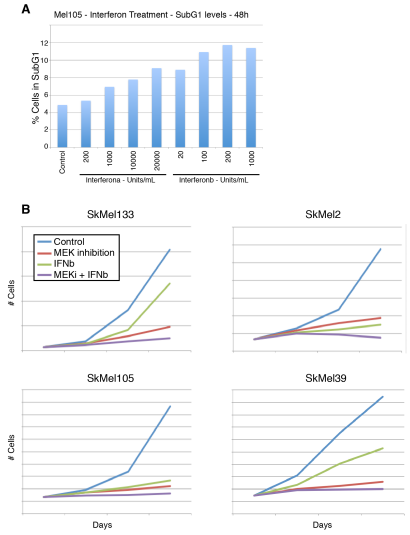


Figure IV-12 – Effects of IFN treatment

A. Dose-dependent response to IFNα and β. The cytotoxicity of IFN was assessed in high pSTAT1 cell line, 48 hours after treatment using SubG1 percentage. IFNα has a weaker cytotoxic effect than IFNβ, and both show dose-dependent effects. 1000Units/mL of IFNβ was used for all experiments in this manuscript. **B.** Growth curves of 2 low- (top) and 2 high- (bottom) pSTAT1 cell lines with MEK inhibition, IFNβ or both.

Transcriptional response to IFN is similar in all cell lines

My data demonstrated that basal activation level of the interferon pathway predicts the cytotoxic response to MEK inhibition, and to its combination with IFNα/β. We hypothesized that differences in the response of the interferon pathway in high- and low-pSTAT1 cell lines to IFN contribute to this phenotype. I therefore characterized the signaling and transcriptional responses to IFNβ and MEK inhibition, aiming to identify the components that contribute to the lack of cytotoxic response in high-pSTAT1 cell lines, and to the synergistic effect of IFNβ and MEK inhibition.

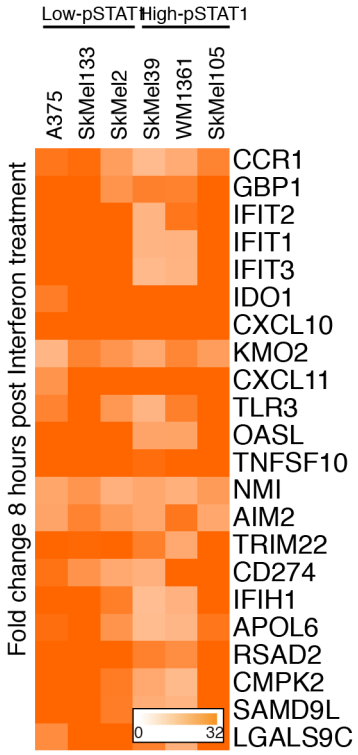


Figure IV-13 - Transcriptional response to IFNβ
22 genes with the highest fold change following IFNβ treatment. The transcriptional response is similar in all cell lines, both with low- and high- basal activation of the pathway. Notably, the fold change of several genes reaches 100 fold, just 8 hours after treatment.

Western blots show that activation of STAT1 by IFNβ is identical, in both timing and extent, when comparing a low-pSTAT1 cell line to a high-pSTAT1 cell line (figure IV-14A). IFNβ treatment quickly elevates pSTAT1 levels in both cell lines. Additionally, both cell lines activate the interferon transcription program, as assessed by levels of Interferon Response Factor 1 (IRF1) [228]. Moreover, inhibition of MEK does not alter the timing or extent of the IFNβ response (figure IV-14A).

To search for more global regulatory differences in the interferon response, I measured gene expression levels 8 hours after treatment with PD325901, IFNβ or their combination in three low- and three high-pSTAT1 cell lines. All cell lines show a dramatic increase (up to 100 fold) in the expression of interferon targets following IFNβ treatment, confirming that the interferon response pathway is present and active in both contexts (figure IV-13A). Furthermore, no significant differences in the transcriptional response following IFNβ treatment between the low- and high-

pSTAT1 cells become apparent after 8 hours of treatment. Additionally, MEK inhibition does not alter the IFNβ response, and does not synergize with interferon to induce transcription of any other genes (see materials and methods).

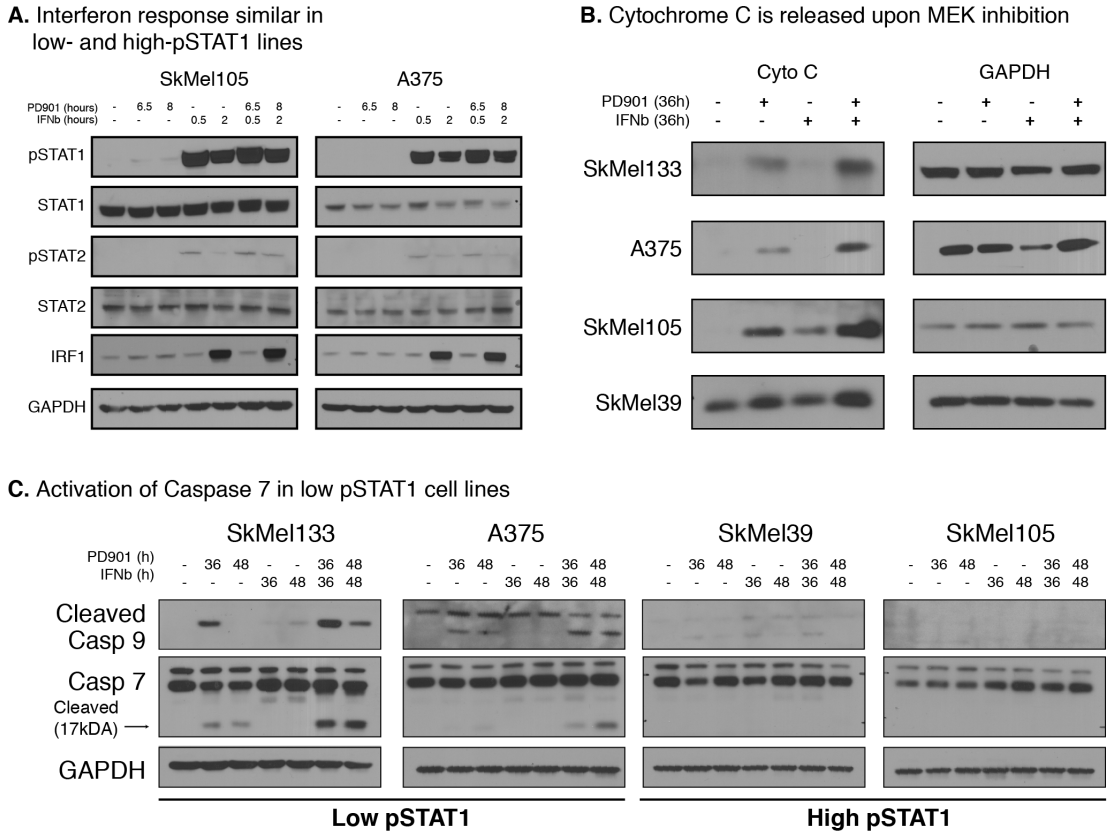


Figure IV-14 - Elucidating the synergistic response of IFNβ and MEKi
A. Response to IFNβ, as measured by pSTAT1 and IRF levels, are similar in both high- and low- pSTAT1 cell lines MEK inhibition doesn't alter the response (for transcriptional response see figure IV-13). Notably, basal activity level of the pathway in high-pSTAT1 cell lines is much lower than the induction in pathway activity after IFNβ treatment. **B**. MEKi activates the intrinsic apoptotic pathway by cytochrome C release from the mitochondria, approx. 36 hours after treatment. IFNβ enhances the response in all cell lines, including the high-pSTAT1 resistant cell lines. **C.** Caspase 7 and 9, but not Caspase 3 (figure IV-15), are cleaved and activated following MEK inhibition in low pSTAT1 cell lines only. IFNβ enhances MEKi's effect, but fails to activate the pathway by itself. Both caspases are not cleaved in high-pSTAT1 cell lines, explaining their resistance to treatment.

These data indicate that the differences in the phenotypic response are not due to the basal activation level of the interferon pathway. The results show that the immediate transcriptional response to IFNβ is not different between high- and low- pSTAT1 cell lines, and therefore fails to explain the synergistic effect of MEK inhibition and IFNβ, and the lack of cytotoxic response in high pSTAT1 cell-lines.

High-pSTAT1 cell lines fail to activate the caspase pathway

As the transcriptional response to IFNβ fails to explain the differences in the cytotoxic response between low- and high-pSTAT1 cell lines, I characterized the apoptotic pathway directly.

The intrinsic apoptotic pathway is initiated by the release of cytochrome C (CytoC) from the mitochondria, which together with Apaf-1, cleaves and activates initiator and executioner caspases [229]. I found that inhibition of MEK is sufficient to induce release of CytoC in all cell lines. Furthermore, co-treatment with IFNβ synergizes with MEK inhibition and increases cytoplasmic CytoC levels (figure IV-14B). However, although MEK inhibition initiates the intrinsic pathway in high-pSTAT1 cell lines, and this response is enhanced by IFN, these cell lines fail to undergo apoptosis.

CytoC release leads to apoptosis by activation of the caspase pathway. I found that caspase 9, an initiator caspase, and caspase 7, an executioner caspase, but not caspase 3, are cleaved following the release of CytoC by MEK inhibition (figure IV-14C and IV-15) in low-pSTAT1 cell lines only. Combinatorial treatment leads to a stronger and faster activation of these two caspases, but IFNβ treatment alone does not activate the caspase pathway (figure IV-14C). Importantly, caspases 9 and 7 are not cleaved in high-pSTAT1 cell lines, although CytoC is released. This lack of activation explains their cytotoxic resistance to both MEK inhibition and its combination with IFN.

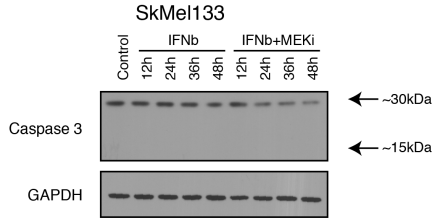


Figure IV-15 – Caspase 3
Caspase 3, an initiator, is not activated by IFNβ and its combination with MEK inhibition.

Deletion of interferon locus correlates with cytotoxic response

I showed that basal activity of the interferon pathway predicts the cytotoxic response to MEK inhibition and its combination with IFNα/β. Therefore, I sought to identify genetic lesions that could be responsible for the differential basal activation of this pathway.

Using The Cancer Genome Atlas (TCGA) melanoma dataset, I associated STAT1 pathway activity levels with genetic aberrations. To infer pathway activity, we used the genes in the STAT1 cluster identified by COSPER, which reflect pSTAT1 levels, as demonstrated by Western blot (figure IV-11A,B). Those genes are also a highly correlated in the TCGA patient derived dataset (figure IV-16A), which allows me to infer pSTAT1 activity in the TCGA tumors. This extended patient-derived dataset enables a genome-wide search for loci whose copy number levels are associated with STAT1 activity (see materials and methods).

The copy number aberration most associated with the STAT1 gene signature is a deletion of the interferon locus (qvalue=10^{-4}), located in chromosome 9p22. The locus contains a cluster of 26 interferon genes (figure IV-16B). Deletion of the locus corresponds to low basal activity of the interferon pathway. My cell line panel confirms this association - most cell lines with low pathway activity have 0 or 1 copies of the 9p22 locus, while all cell lines with high activity have 2 or 3 copies (figure IV-16C, materials and methods).

Interestingly, the interferon gene cluster on locus 9p22 is only 0.5Mbs downstream of p16 (CDKN2A) (figure IV-16B), a known tumor suppressor gene deleted in roughly 60% of melanoma tumors [230]. Deletion of both *p16* and the interferon locus was previously reported [231], but as research focused on the role p16 in cancer, deletion of the interferon locus was viewed as a passenger mutation. However, copy number data show that both events are independent, and cell lines can lose both copies of *p16* but retain both copies of interferon locus. Taken together, these results suggest that deletion of the interferon locus has important consequences on cellular phenotype, and its role is independent of *p16*.

High basal pSTAT1 activity is caused by an autocrine loop

Copy number data show that cell lines with low basal activity of the interferon pathway have fewer copies, on average, of the interferon genes. IFNβ treatment elevates pSTAT1 levels and its downstream targets to similar levels in all cell lines. I therefore hypothesized that the basal activation of STAT1 and interferon pathway in cell lines and tumors might be due to expression of the genes in the interferon locus, which act in an autocrine loop to activate the pathway.

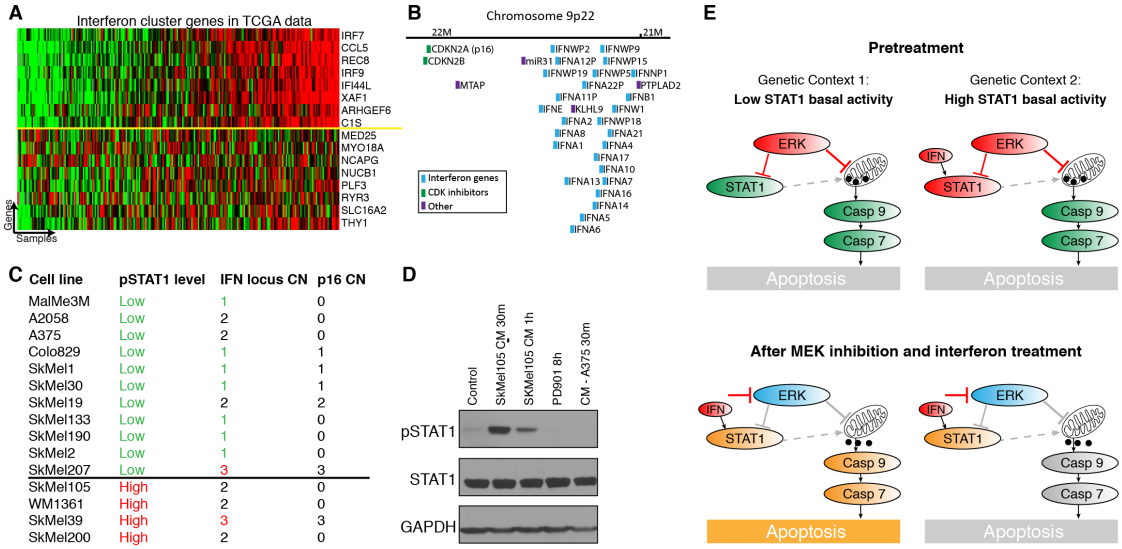


Figure IV-16 - Deletion of interferon locus and IFN expression levels explain the two interferon-pathway states and predicts drug response

A. The interferon gene cluster identified by COSPER is highly correlated in the TCGA melanoma expression data set. This allows me to infer pathway activity in the TCGA tumors and associate it with DNA aberrations. Genes above the yellow line were used for association with DNA copy number. **B.** The interferon locus contains 17 interferon genes, and is only 0.5Mb downstream of CDKN2A (p16), a known melanoma tumor suppressor. **C**. Interferon locus copy number is also correlated with pathway activity in our 14 cell line panel. p16 however, only 0.5Mb upstream, is not, suggesting that interferon deletion and p16 deletion are two independent events. SkMel200, a high-pSTAT1 cell line, was added for purposes of CNV analysis. **D**. Conditioned media experiment demonstrating that cytokines released to the media elevate pSTAT1 levels. Media taken from SkMel105, a high pSTAT1 cell line, quickly elevates pSTAT1 levels when applied to a low pSTAT1 cell line. Several IFN genes are upregulated in high-pSTAT1 cell lines (figure IV-14A). **E**. A cartoon depicting the two network states, before and after MEKi and IFN treatment. Inhibition of MEK leads to cytochrome C release in both cellular contexts, and IFN treatment enhances the response. However, caspase 9 is cleaved and activated only in low pSTAT1 cell lines.

Conditioned-media experiments confirm this hypothesis. In these experiments, conditioned media from a high-pathway-activity cell line activates the interferon pathway in low pathway-activity cell line (figure IV-16D). These results demonstrate that cytokines in the media, presumably IFN, lead to the high basal interferon pathway activity in cell lines without a deletion of the interferon locus. Moreover, several interferon genes, including IFNb, which is located in the

deleted locus, are correlated with pathway activity and overexpressed in high-pSTAT1 cell lines (figure IV-17).

To summarize, my results show that cell lines with fewer copies of the interferon locus and without expression of the interferon genes are sensitive to the cytotoxic effects of MEK inhibition. Furthermore, IFNα/β enhances this cytotoxic response via an increase in CytoC release from the mitochondria. However, cell lines that retain both copies of the interferon locus and have high basal activity of the interferon pathway are resistant to MEK inhibition and its combination with IFNα/β. Although MEK inhibition initiates the apoptotic pathway in these cell lines, apoptosis is averted due to an impairment of the caspase activation mechanism. It is therefore possible that interferon pathway activity, interferon expression levels and/or interferon locus copy number can be used as a biomarker for treatment by MAPK pathway inhibition, and its combination with IFNα/β (figure IV-16E).

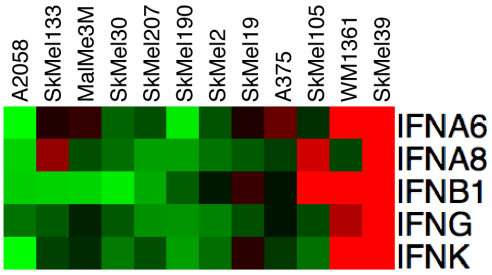


Figure IV-17 – Expression of interferon genes
IFN genes with a significant differential expression between low- and high- pSTAT1 cell lines. IFNA6, IFNA8 and IFNB1 are located in the interferon locus.

Materials and Methods

Cell Culture and drug treatment

Cell lines were obtained from N. Rosen (Memorial Sloan-Kettering Cancer Center), except for Colo829 and A2058 that were purchased from ATCC. All cell lines were maintained in RPMI 1640 (Invitrogen 21870-092), supplemented with 2 mM glutamine, 50 units/mL penicillin, 50 units/mL streptomycin, and 10% FBS (Omega Scientific), and incubated at 37 °C in 5% CO2.

Samples for protein and gene expression analysis were plated at 60-80% confluency and incubated for 20-24h. Then treated with PD0325901 (50nM), Interferon alpha (20000U/mL, R&D 11100) or Interferon beta (1000U/mL, R&D 11415). Control samples were collected untreated at time of treatment.

Protein levels

Samples for protein analysis were lysed using RIPA buffer. Protein concentration was assessed using BCA staining. Samples were then normalized to a fixed concentration and mixed with a 5x glycerol/SDS/DTT loading buffer. Lysates were run on gradient (4-12%) Bis-Tris gels.

Cellular fractionation for nucleus isolation to assess NF-κB activity was performed using Subcellular Protein Fractionation Kit by Thermo-Fisher.

Cytochrome C release

Protocol for Cytochrome C release is taken, as is, from Majewski et al 2004. It is brought here for convenience:

Lysis buffer: 20 mM Hepes-KOH, [pH 7.5], 210 mM sucrose, and 70 mM mannitol; 1.5 mM MgCl2, 10 mM KCl2, protease inhibitor, and 1 mg digitonin/1mL lysis buffer.

Cells are trypsinized, collected and spun down in 4C. They are then washed with PBS and spun down again. It is critical that cell pellets will be lysed immediately without freezing.

Cells are gently suspended, without vortexing, in lysis buffer. Roughly double the cell pellet volume is used. They are incubated in 25C for 3-10min, depending all cell line. Spun down at 4C for 20 minutes at highest speed. Supernatant contains cytoplasmic fraction.

Protein concentration was assessed using BCA.

Growth curves and Apoptosis levels

For growth curve measurement, 50K cells were plated in 6-well plates with 2mL of growth media. Cells were counted every 24h following treatment using a cell counter (Coulter Z1), in triplicates.

Apoptosis was assessed by TUNEL staining. Cells were plated in 6-well plates at 200K cells/well. 24h after plating cells were treated with PD325901, and both floating and adherent cells were collected 72h after treatment. TUNEL was performed using Invitrogen BrdU TUNEL kit.

Growth rate phenotype for the STAT3/NF-κB analysis was calculated by dividing number of cells after 4 days of treatment by the number in day 0.

Gene expression and microarrays

Samples for microarrays were harvested 8h post treatment. RNA was extracted using a Qiagen RNeasy kit, and labeled using Agilent's one-color labeling protocol. Labeled cRNA was hybridized to Agilent's 8x60 human gene expression arrays. MEK inhibition and basal state expression levels were measured in biological duplicates. Data normalization is described in supplementary material. Genatomy was used for data visualization and enrichment analysis [158].

I used Agilent's 1M SurePrint CGH arrays to assess copy number. DNA was extracted using Qiagen's DNeasy kit and labeled and hybridized according to Agilent's protocol.

All microarray data are available on GEO under accession number GSE51115.

Microarray preprocessing

Agilent one-color human mRNA expression 8x60 arrays were used to assess expression levels. Biological duplicates of control and MEK inhibition (MEKi) samples were used (expect for Colo829 and SkMel28 that were added to the panel after the first batch). Samples for the IFNβ microarrays were collected 8h after treatment (with IFNβ, PD901 or both), and a single sample was used for each.

Agilent's software was used to assess raw signal intensity. Preprocessing of both the MEKi panel and the IFN experiment was similar. Each of the 3 batches were processed independently - MEKi panel 1, MEKi panel 2 and the IFN panel.

Preprocessing consists of 3 steps – probe filtering, data normalization and probe averaging.

Probe filtering

Log2 values were used from this point on. Probes were filtered based on their values. Probes with low or high levels in more than 20% of samples were removed. This was done to remove noisy and saturated probes. The lower and upper thresholds were different in different batches, depending on labeling, hybridization and scan levels:

Batch	Lower threshold	Upper threshold
MEKi panel 1	6	16
MEKi panel 2	7	18
IFN panel	7	17.5

Additionally, the Agilent probe flags were used to filter probes by a similar method: probes flagged in more than 20% of samples were removed. Flags that were used: will_above_bg, is_saturated, is_feat_non_uniform, is_feat_popn.

A “rescue” step was used to return probes representing genes that no probe was left to represent them. Probes representing the same gene with a high correlation (Pearson >0.75) were rescued. Additionally, probes with high SD (>3) were also rescued.

Data normalization

The 75th percentile of all samples was set to the average 75% by multiplying the values by a constant.

Probes that measure the level of the same gene were averaged or filtered out.

If the average Pearson correlation between all probes is > .75, probes are averaged. If it is lower, the probe with the lowest correlation is removed. Process repeats till probes are averaged or only one probe is left.

Merging duplicates

Baseline expression levels are mean-normalized at the gene level. Fold change is calculated against the control (baseline expression) of the cell line. Data from the two MEKi panels are combined at this point by averaging the baseline expression and fold change data.

TCGA data analysis

TCGA expression and CGH data were downloaded from the TCGA website. Genes for the STAT1 gene signature were a subset of COSPER's STAT1 signature. All genes with a Pearson r^2 > 0.5 with at least 3 additional genes were included. Association with copy number was performed using Pearson correlation between the mean of the gene signature and copy number levels of each gene. Pearson's pvalues were corrected by FDR [181].

Comparison of BRAF and MEK inhibition - PLX4720 vs. PD901

I used PD901 to inhibit the MAPK pathway, and not the more clinically used PLX4032 BRAF-V600E inhibitor to allow a direct comparison of BRAF and NRAS mutant cell lines. To ensure the short-term drug effects are similar, we compared the transcriptional response of MalMe3M, a BRAF-V600E cell line, following PD901 or PLX4032 treatment. I assessed expression fold change at 1 hour, 2, 4, and 8 hours following treatment using Illumina HumanHT-12 microarrays.

Preprocessing

Illumina's probe pvalues were used to filter out probes. Probes with p-value>0.05 in more than half of the samples were removed. Then microarrays were normalized according to their 75% percentile values. The 2 control array were averaged, and treated samples were compared to the averaged control to assess fold change.

Results

MEKi and BRAFi are remarkably the same at all time points. Although some probes were noisy, resulting in minor difference between treatments, no gene had a difference greater than 0.5 fold (on a log2 scale) between treatments at all time points. Only 6 probes, out of 16000, had a difference of more than 1 fold at 8 hour time point (figure IV-18). None of them had such difference at 4 hours, suggesting that the difference arises from measurement noise.

I conclude that there is no difference in the short-time transcriptional response between treatments in this cell lines.

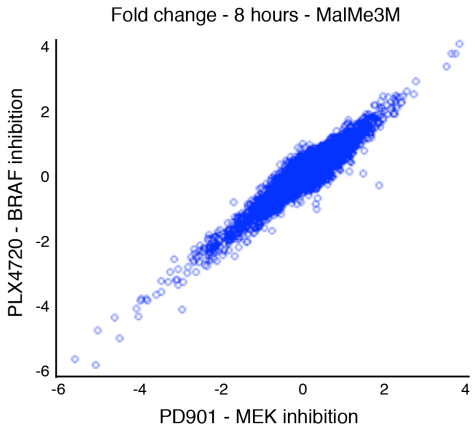


Figure IV-18 – Comparison of MEKi and BRAFi
Comparison of MEK and BRAF inhibitors in a BRAF-V600E cell line shows an almost identical transcriptional response. Scatter plot shows fold change of all genes with a MEK inhibitor (x-axis) compared with a BRAF inhibitor (y-axis). Almost all genes fall directly on the diagonal.

PD901 and IFNβ microarray results

Data Preprocessing

Six cell lines were chosen for analysis. 3 are low-pSTAT1 – A375, SkMel133 and SkMel2, and 3 high-pSTAT1 – SkMel105, SkMel39 and WM1361. They were treated with 50nM PD901, 1000U/mL IFNβ or their combination. Samples were collected 8 hours after treatment, control samples were collected at 0h. RNA extraction, labeling and hybridization were conducted as described under methods. Agilent human 8x60 gene expression arrays were used.

Raw data normalization and filtering were conducted as described above, with a low threshold of 7, and an upper threshold of 17.5.

IFN response in high- vs. low- pSTAT1 cell lines

The IFN response includes dozens of genes with a dramatic induction in gene expression, of up to 500 fold, in all 6 cell lines (figure IV-13).

There is, however, a difference in the extent of change in high- vs. low- pSTAT1 cell lines, that can be attributed to the different basal expression level of those genes. The maximum level of expression seems to be similar in all cell lines, but high pSTAT1 cell lines have a higher basal activity and therefore the fold change is lower.

In order to compare the activation of the pathway between the two cell line groups, it is better to use the final expression level, i.e. the basal expression+fold change. However, such comparison reveals the expression of no genes is statistically significant different between high- and low-pSTAT1 cell lines (using t-test and FDR correction).

I therefore determine that there is no difference in the response to IFNβ between high- and low-pSTAT1 cell lines.

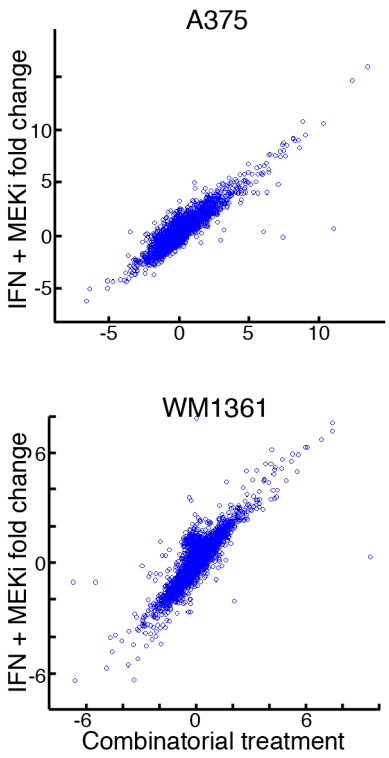


Figure IV-19 – Transcriptional response isn't synergistic
Lack of synergistic and additive effects of MEK inhibition and IFNβ. Scatter plots show the fold change of all genes with a combination of MEK inhibition and IFN (x-axis) and the sum of fold changes with each treatment alone. Significant deviations from the diagonal represent synergism between drugs. Only one gene, CCL4, deviates from the diagonal in all 6 cell lines

Combinatorial treatment and synergy

To test whether the MEK inhibition and IFNβ synergize at the level of gene expression, I compared the fold change of the dual treatment with that of MEKi+IFNβ as single agents. Over all, those responses are very similar (figure IV-19).

If no synergy exists, the values of Both-(MEKi+IFNβ) should be close to 0. Only one gene significantly deviates from 0 in all 6 cell lines. The gene is CCL4, and it is induced both by MEKi and IFNβ treatment as single agents, but a combinatorial treatment isn't additive.

I couldn't identify any other genes that show a synergetic response in all 6 cell lines, or separately in low- or high-pSTAT1 lines (we defined synergy is the equation above >1 or <-1).

Perturbation data allows better cluster identification

Combining pre- and post-inhibition data facilitates the identification of context-specific regulation and differential activation of pathways, while pre-inhibition data alone fall short due to lower specificity and much higher rate of false positive.

For example, when running the first step of COSPER on pre-inhibition data alone, the STAT3-NFkB cluster contains 766 genes, compared with 28 genes when using both datasets. While the smaller cluster is enriched for STAT3- and KFkB- related terms, the larger pre-inhibition-data cluster is enriched for general terms such as "extracellular region" and "plasma membrane".

The combination of pre- and post-inhibition data, therefore, provides specificity and limits the cluster genes to only genes directly regulated by MAPK, while also provides the context of regulation.

Computational Methods

Tippi – Target Identification using Pre- and Post-Inhibition data

Tippi identifies target genes of a pathway (in this case MAPK) using gene expression data of pre- and post- perturbed cells. Most target identification methods are based on the fold change of a gene post-inhibition of the pathway, and require a gene to have a significant change in X or more samples. However, Tippi's definition for a target is a gene with a negative correlation between the expression changes following a perturbation (i.e. pathway inhibition) and base line expression levels (steady state, pre-perturbation).

Tippi scores genes using Pearson correlation and then uses permutation testing to assess the significance. Tippi uses the permutation p-value to construct the final target list. To compute Pearson correlation, I use steady state levels correlated against the fold change of the gene across all cell lines. However, such score is prone to spurious associations due to outliers. To overcome this problem, I take a leave-one-out approach. In the leave-one-one process, I remove one sample at a time and calculate the Pearson r-value for a gene. The leave-one-out score, or the Tippi score, is the maximum r-value of the process.

To assess the significance of the leave-one-out score and further filter false-positives, I use permutations. Samples are permuted in one of the datasets (steady state or fold change), and calculate the leave-one-out score. The process is repeated 1000 times, and the gene's p-value is the number of permutations with a leave-one-out score better (lower) than the original (un-permutated) score.

I repeated the process for the two biological replicates independently. One datasets contains all 14 cell lines, and the other includes 12 of those (SkMel28 and Colo829 are not included). The final list includes all genes with a p-value<0.05 in both datasets. 1985 genes pass this threshold in the first batch, and 2012 pass it in the second batch. With an overlap of 793 genes (Hypergeometric p-value of 10^{-173} for the overlap).

COSPER - Context-Specific Regulation

COSPER – COntext SPEcific Regulation – is designed to identify genes that are directly regulated by the MAPK pathway (or any other perturbed pathway) in only a subset of cell lines. It is based on the assumption that genes under the direct control of a pathway are correlated before pathway inhibition and show a correlated expression change after pathway inhibition. Since we are looking for genes under the control of the pathway in only a subset of cell lines, we expect expression changes in only these cell lines.

COSPER uses pre-perturbation data to limit the search for genes under direct regulation of the perturbed pathway. After inhibition of a key signaling pathway such as MAPK, cellular events, such as metabolism, cell cycle and apoptosis, lead to expression changes of thousands of genes. Although the expression of those genes changes after MAPK inhibition, they are not directly regulated by MAPK. However, genes under the direct control of MAPK pathway depend on its activation levels both before and after inhibition of the pathway. For example, HEY1 (figure IV-6A) is under the control of MAPK in only a subset of cell lines. In HEY1 case, it is overexpressed by MAPK in cell lines with high MITF levels. Therefore, only in MITF-high cell lines, HEY1 expression levels decrease after MEK inhibition. Both pre- and post-inhibition expression levels are needed in order to determine this relationship.

COSPER is therefore designed to find genes with context-specific regulation patterns (figure IV-6B). It is consists of 3 major steps:

1. Gene level – identify binary splits with high scores for both baseline expression and fold change and construct clusters.
2. Merge related clusters – allows removal of spurious correlations and averaging the noise caused due to the small sample size.
3. Add high scoring genes to the remaining clusters

A detailed description of each of the steps follows the section on the NormalGamma score.

NormalGamma score

The algorithm is based on the NormalGamma score [169,232]. The NormalGamma is a Bayesian score that takes variance, mean and number of data points into account. It gives a higher score to a data matrix with low variance.

I use this score since we are looking to reduce the variance of the samples. COSPER searches for genes that behave similarly in a subset of samples. For example, I look for a subset of samples where a predefined set of genes is up-regulated, compared with the rest of the samples where the genes are not under pathway control. Mathematically, this problem can be viewed as a subset of samples where the data have a lower variance compared with the variance of all samples combined. The NormalGamma score is driven mainly by data variance and is thus suitable for our needs.

The score:

$N = size(data)$

$\beta = \max\left(1, \frac{\lambda(\alpha-2)}{\lambda+1}\right)$

$\beta_{plus} = \beta + \frac{Var(data)N}{2} + \frac{N\lambda|data|^2}{2(N+\lambda)}$

$\alpha_{plus} = \alpha + \frac{N}{2}$

$NormalGamma(data, \lambda, \alpha) = -N * \ln\left(\sqrt{2\pi}\right) + \frac{\ln\left(\frac{\lambda}{\lambda+N}\right)}{2} + \ln\left(\Gamma\left(\alpha_{plus}\right)\right) - \ln\left(\Gamma(\alpha)\right) + \alpha\ln(\beta) - \alpha_{plus}\ln\left(\beta_{plus}\right)$

The score used to assess the quality of the split is:

NormalGamma (right samples)+ NormalGamma (left samples)- NormalGamma(all samples)

Step 1: Creating clusters

First, gene expression is normalized. Basal expression levels of each gene are set to have μ=0 and σ^2=1. Fold change for each gene is standardized only (σ^2=1).

Next, clusters are built bottom-up – genes are assigned to “splits”, and a split with more than one gene assigned to it is considered a cluster. However, in order to filter out spurious associations I only consider clusters with 5 or more genes. All genes are tested across all valid

binary splits. A valid split assigns at least 2 samples to each sample group. The test is based on permutations and the NormalGamma score.

A gene is assigned to a split if its NormalGamma scores (as defined in the previous section) in both the baseline expression and fold change are better than 99% of the split permutations (pvalue<0.01).Additionally, in order to keep the best split-gene pairs only, an additional threshold is used:

NormalGamma (right)+ NormalGamma (left)- NormalGamma(all samples)>0

To determine whether clusters with more than 5 genes can be constructed by chance, I permuted the samples in the fold change expression data and performed this step on the permutated data. No clusters with 5 or more genes were constructed. Hence, the resulting clusters represent biological phenomenon.

Step 2: Merging clusters

A gene assigned to a split is very likely to be assigned to similar splits. A similar split might have one or more samples switching "sides" (figure IV-6C,D). Each split has 13 similar splits with a distance=1, where one sample has switched sides, and 91 splits with distance=2.

The NormalGamma score is not strong enough to discriminate between the "true" split and neighboring splits, since the distribution of scores is very tight. However, I can assume that a gene is more likely to be assigned to the real biological split, and less likely to be associated with a split with a distance>0 from the real split. We also work under the assumption that a true biological "context" is likely to influence many genes, and therefore larger clusters are more biologically relevant.

I use these two assumptions in order to identify the real gene-split associations and remove irrelevant clusters.

The cluster merging algorithm is a iterative process. Each cycle identifies the largest cluster, its genes are removed from all its neighboring clusters, and the process iterates till no more clusters can be identified.

The steps are:

1. Each cluster is scored based on its overlap with its neighbors:

$$Score(cluster_x) = \sum_{i \text{ where } Distance(Split_X, Split_i) \leq d} \#(Genes_{Cluster_x} \cap Genes_{Cluster_i})$$

   we used d=2.
2. We then choose the largest cluster, and remove its genes from all clusters with a distance<=d.

To save computing time, only clusters that enter the algorithm with 5 or more genes are allowed to be selected.

Step 3: Adding genes to remaining clusters

In the last step, after filtering most clusters out, I allow genes from neighboring clusters to be added back to clusters. We found this step to be necessary due to the small sample size, the overall small distance between clusters, the relatively high noise of gene expression data, and the inability of the NormalGamma score to discriminate between similar splits.

Genes belonging to clusters in a distance<=d of a specific cluster, and with a pvalue<0.01 are added to this cluster.

MITF binding site analysis

To assess frequency of MITF binding site in gene promoters we used the motif CACATG, known to be a target sequence of MITF. Gene promoters were defined as 5000bp upstream of their transcription start site, or up to the closest upstream gene, whichever is shorter. For each gene, number of binding motif in its promoter sequence was noted.

To assess the significance of number of motif occurrences, we used the binomial distribution. For each one of the two clusters, MITF-M and MITF-expression, we counted total number of motif occurrences in all the cluster genes. For simplicity, the expected probability of the motif to randomly appear in a DNA sequence is 2*1/46 (6 is the length of the motif, and 2 represent the two strands).

The pvalue of X occurrences is the probability of randomly observing X or more occurrences in a random sequence, or 1-BINOMIAL_CDF(X, N, p), where N is total sequence length and p is 2/46.

For MITF-M cluster, the total promoter sequence is 120735bp, with 83 motif occurrences (59 expected). For MITF-expression cluster, the total promoter sequence is 183399bp, with 86 occurrences (89 expected).

Discussion

Contemporary cancer drug development focuses on targeting recurring oncogenic events, such as gene amplification and overexpression (HER2) or activation (BRAF). This approach is based on the principle of oncogene addiction. However, the underlying assumption is that both the network structure and the downstream targets of the oncogenes are the same in all tumors. Taken further, drug combinations are also currently suggested based on the principle of similar network structure and pathway dependencies in tumors harboring a specific oncogenic mutation.

However, my analysis of downstream targets of MAPK in MAPK-activated melanomas reveals tremendous differences in underlying network structure between tumors. Although I analyzed the transcriptional output of MEK inhibition only in melanoma cell lines with MAPK activating mutations (BRAF or NRAS), each cell line had a unique transcriptional signature. Moreover, a vast majority of downstream targets of the MAPK pathway are *context-specific* – under the control of the pathway in a subset of cell lines. I showed that these differences could predict the phenotypic heterogeneity observed *in vitro.*

To detect context-specific targets using pre- and post-inhibition expression data, I developed and employed two algorithms, Tippi and COSPER. With Tippi I showed that the expression changes of hundreds of downstream genes following pathway inhibition are proportional to the expression levels at steady state. I then used COSPER's clusters to identify the pathways that control the expression of the context-specific targets. For example, STAT3 and NF-kB were shown to have two activation states that influence MAPK targets. Moreover, I found that the activation states of these pathways are correlated with the growth rate under MEK inhibition.

Tippi's and COSPER's results emphasize the importance of post-perturbation data. Even with a small sample size of only 14 cell lines, pre- and post- perturbation expression data empowers the discovery of dependencies and interactions between pathways. Larger datasets of pre- and post-inhibition expression data can help identify additional context-specific interactions, which are masked by the substantial influence of MITF status on gene expression in melanoma. Moreover, interactions between pathways can inform me about possible interactions between drugs. COSPER's output provided me insights on a possible interaction between MEK inhibition and IFN

treatment, two approved treatments for melanoma. The experimental validation uncovered two key findings: first, IFNα/β enhances the cytotoxic response of MEK inhibition; second, cell lines with high basal activity of the interferon pathway exhibit much lower cytotoxicity under MEK inhibition. I was also able to identify that a deletion of the interferon locus is correlated with the basal activity level of the interferon pathway. However, my results indicate that the basal activity level is not the mechanism for the sensitivity and resistance to IFNα/β and MEK inhibition. Instead, I found that an impairment of the caspase activation mechanism leads to the cytotoxic resistance.

I also demonstrated that MEK inhibition leads to, and IFNβ increases, the release of CytoC from the mitochondria in all cell lines, regardless of their interferon-pathway basal activity level. Following CytoC release, caspases 9 and 7 are activated only in cell lines with low interferon pathway activity. Cell lines with high basal pathway activity, however, do not cleave and activate caspase 9 following MEK inhibition, and apoptosis is averted.

Interferon pathway activity was previously linked to drug response. Weichselbaum et al. [113] have shown that breast tumors with high basal activity of the pathway are more resistant to chemotherapy and radiation. TCGA data show that a lower basal activity of the interferon pathway in breast cancer is associated with a deletion of IRF1, Interferon Response Factor 1, a necessary protein for interferon-induced death[233]. Taken together, we postulate that constitutive exposure to IFN is adverse to cancer cells, and they overcome it by either deactivation of the interferon pathway, or by an impairment of the apoptotic pathway.

My findings on IFNα/β and MEK inhibition could have important clinical implications. First, both IFNα/β and MEK/BRAF inhibition are approved treatments in melanoma, and a combination might be beneficial to patients. Moreover, this combination might specifically benefit NRAS melanoma patients, who are treated with low doses of the more toxic MEK inhibitors, compared with the BRAF-mutant specific drugs used in BRAF-patients. Second, the impairment in the caspase pathway might be clinically important in melanoma and other cancer types. Finally, it is possible that interferon pathway activity and/or the interferon locus can be used as a biomarker in cancer treatment.

The current paradigm in MAPK pathway inhibition aims at a complete blocking of pro-survival signaling. Suggested combinatorial treatments include combination of MAPK pathway inhibitors (such as RAF and MEK inhibitors [234]), or combinations that prevent the feedback activation of RTKs [234]. However, examination of the pathway interactions and analysis of transcriptional response following MEK inhibition identified a drug combination that takes a different approach. Instead of exerting all effort on shutting down MAPK signaling, I found that IFNβ, which works via a different signaling pathway, enhances the cytotoxicity of MEK inhibition. I believe that a COSPER-like analysis of pre- and post-perturbation data could reveal additional combinations of drugs that synergistically work on different pathways in other cancer types.

To summarize, my work demonstrates that tumor networks are more complex and varied than previously appreciated, even within a subtype of cancer that shares key oncogenic mutations. Although only MAPK-activated melanoma cell lines were examined, these were found to be heterogeneous and immensely varied. Moreover, while all BRAF-mutant tumors are grouped together and treated similarly in the clinic, the targets and pathways regulated by BRAF in different cell lines are vastly different.

The full scale of these differences is only revealed when examining a perturbed network, which highlights the importance of post-inhibition data, compared with steady-state data only. I believe that my research has only scratched the surface, and future studies with larger cohort size should be conducted, as my data demonstrate the value of system-wide perturbation analysis of tumors in the era of personalized medicine.

Discussion

Computational biology addresses a broad spectrum of questions and tasks, each of which uses different types of data, has different assumptions on the underlying model, and aims to answer different questions using different computational models.

My work has touched several different tasks. I analyzed both yeast and cancer data using various data derived from various technologies – from high throughput genomic data, through growth curves and ELISA, to Westerns and qPCR. I addressed research problems and developed computational methods for association studies, driver identification and the underlying molecular mechanisms of resistance to drugs. My work spans various biological and computational fields, but I find that regardless of the organism I worked on and the biological problem I addressed, my research was guided by similar concepts.

All my computational models account for context-specificity of the biological system. The computational methods were simple and had few assumptions about the underlying biology. And although I used high throughput and complex datasets, I attempted to investigate well-defined biological phenomena. The reasons I adhered to these guidelines are described in this discussion.

Defining a phenotype

Addressing well-defined biological phenomena should be at the foundation of every research project, including computational modeling. In many cases, especially in the field of systems biology, a computational method is designed to identify features (genomic, genetic, etc.) that predict or explain a phenotype. A pivotal decision in project design is therefore the definition of the phenotype itself. Since many computational methods use a predictive model that assumes that one mechanism underlies the phenotype in all individuals, a good choice of phenotype would be one that is likely to be a result of one underlying mechanism.

The project modeling the phenotypic heterogeneity of MAPK inhibition in melanoma exemplifies the importance of selecting an appropriate phenotype. There are several ways to assess the outcome of a treatment in an *in vitro* setting. The most popular phenotypes, including

IC50 and growth curves, are based on number of cells, i.e. the quantitative phenotype is based on the number of cells following a treatment compared to the number of cells before or with no treatment[46]. However, cell number following treatment is determined by several distinct cellular processes, including cell cycle and death. Each of these processes is likely to be controlled by a different underlying mechanism, or pathway. Computational methods designed to identifying one underlying mechanism for phenotypes that are a result of several underlying mechanisms, such as cell-number based phenotypes, are unlikely to find the true underlying mechanisms.

In the case of drug sensitivity, especially in cases with small sample size, a better approach would be to model the various effects of the drug independently of others, e.g. cytotoxic, cytostatic, senescence. In my study I found that different cell lines respond in very different ways to MAPK inhibition. While some cell lines undergo apoptosis while continuing to proliferate, others fully arrest, but don't die. Therefore, studies aimed to identify the mechanism for one such phenotype are more likely to succeed. It is of course possible that certain phenotypes, such as apoptosis, are a result of different molecular mechanisms in different individuals, and a good computational method should account for such heterogeneity.

One phenotype – One mechanism?

Choosing an appropriate phenotype is only the first step in the design of a computational biology project. The selection of a mathematical method that best models the expected underlying mechanism is also crucial for the success of the project.

In computational biology, the feature space (e.g. genes) is typically much larger than the sample space (e.g. cell lines), especially in genomics based studies. Therefore, the statistical burden of feature selection (e.g. identifying the features that best predict the phenotype) is very challenging. Under these conditions, the mathematical method has many possible solutions, and in order to create a robust method in conditions of vast uncertainty, the methods have to choose models that significantly simplify the problem. One of the most common techniques used to simplify the models is a linearity assumption, which predetermines the relationship between the features and the outcome, forcing it to be linear in nature[45,46]. The linearity assumption greatly

shrinks the solution space, allowing the method to be robust and consistent. However, if the true relationship is far from linear, the method can identify the wrong features.

On top of the linearity itself, linear models also impose an extremely strong assumption on the feature's influence on the phenotype – persistence. Simply put, *persistence* means that the same genomic feature exerts influence all individuals/cases. In such models, the phenotype is the combined result of all features in all individuals, even in cases in which a feature exerts no influence on the phenotype. Due to the small sample size typically used in these studies, *persistence*-based approaches are unlikely to identify the correct features, simply because the model's assumptions are wrong. Although one mechanism can regulate a phenotype in all individuals, my research in several biological systems shows that this is the exception rather than the rule.

To avoid making the linearity and persistence assumptions in the MAPK project, I decided to forgo predictive models altogether and instead model molecular interactions between the targeted pathway (MAPK) and other pathways. I hypothesized that by identifying pathways and processes that are under the control of MAPK, I will be able to learn about the molecular events that lead to the cytotoxic, cytostatic and other possible responses following MAPK inhibition. As the method is not aimed at predicting a phenotype, it doesn't assume that one molecular mechanism underlies a phenotype in all cell lines, and allows for the discovery of context-specific mechanisms.

Implications of Context-Specificity and Heterogeneity

Context-specific models, in which a genomic or a genetic feature exerts influence on the phenotype only in the context of another feature, were a crucial part in my work. Unlike *persistent* models, context-specific models take heterogeneity of the underlying mechanism into account. This allows subsets of cases to be modeled independently of the others, which proved to be invaluable and necessary in the biological systems and questions I investigated.

Both in drug screens and GWAS, in cancer and yeast, context-specific models were at the core of my methods. In GOLPH, context-specific interactions were 4 times more prevalent than linear interactions. In CONEXIC we showed that by accounting for context-specificity our model was able to identify drivers that were missed by other models. COSPER explicitly searches for

context-specific interactions, and identified interactions that underlie resistance to treatment in melanoma. The context-specificity of the MAPK targets in melanoma also exemplified another key phenomenon – the heterogeneity of network structure.

The predominance of network heterogeneity affects a broad range of tasks and tools in computation biology, from the design and application of computational methods, to the post-analysis and interpretation of the results. In most of these tasks, researchers tend to ignore heterogeneity and draw broad and general conclusions based on private cases. For example, many in the field of systems biology, including myself, are inferring "the network" under a transcription factor or a signaling pathway using post-perturbation data. However, a quick literature review shows that in many cases, only one or two models/cell-lines are used[235,236]. In such cases, while the inferred network might be correct, it only reflects a network state in one cell line, under one experimental condition. As my MEK inhibition results have demonstrated, network downstream of MAPK is vastly different in different cell lines. Therefore, these results can't be generalized to be the "global network" in all cell types or all living organisms. These misleading conclusions later hinder the analysis and interpretation of other computational results, by providing wrong lists of "targets" or "gene-sets" that are used for enrichment analysis of clustering results.

In summary, biological systems present vast heterogeneity. The high variability between individuals, or contexts, requires careful design of the computational method, and even more so in the proper use of data from other systems and contexts to interpret the method's results.

The premise of computational biology

Biological systems are of vast complexity. High throughput technologies and statistical methods provide important tools in the analysis and understanding of these systems. However, because of the high complexity of the investigated systems, one has to practice extreme caution when interpreting results of computational models.

Many of the questions addressed using computational methods in the field of systems biology are too complex, currently unsolvable and should not be naively approached with such tools. Theoretical research, of course, is necessary for the evolution of the field. For practical tasks,

however, computational methods must be carefully framed and applied only in situations where the mathematical models match the underlying biological mechanisms, and the data used are appropriate for the questions asked.

Even after choosing appropriate questions, models and data, systems biology tools produce many false-positive results. Therefore, the outcome should not be used as the "true underlying biology", but rather as hypothesis generation to direct and focus future research. For example, CONEXIC aimed at identifying CNV-driven oncogenes by using only CNV and expression data. By ignoring other critical factors such as mutations and epigenetic state, CONEXIC can't explain all the variance in the data. However, the method fits all available data into a very limited conceptual model, inadvertently producing many false-positive predictions. In our research, we used CONEXIC's output to identify two previously unknown drivers, RAB27A and TBC1D16. Our success in identifying those drivers by no means suggests that all other predictions made by CONEXIC are true. We merely used a computational method as a tool to guide and direct our research, and only after a careful examination of the results we were able to select those drivers.

One might refer to this practice as "cherry-picking" - carefully selecting one true result to claim that the rest of output is also true. However, such claims should not, and were not, made. CONEXIC and other computational tools are of great value and strength, but they are not designed to find the entire underlying network. Computational modeling is a tool, not the goal. Its goal is to help researchers direct their efforts towards the more probable answer.

To conclude, while computational tools are powerful and important in biology research, they should be carefully designed and applied, and their results should be interpreted in the scope in which they were applied.

Future Directions

In this dissertation I presented 3 very different studies. Each of these studies examines new computational and conceptual approaches to answer burning biological questions. My work by no means provides a solution to any of these problems. If anything, it provides a glimpse into the landscape of the underlying biology, and explores the possible computational tools that can be used to model this landscape.

GOLPH examined the landscape of genetic interactions in yeast and showed that a large percentage of expression variance can be explained with context-specific interactions. In collaboration with Kreimer and Pe'er[20], I later expanded GOLPH to mammalian data. However, even with GOLPH's context-specific interactions, most of the gene expression and phenotypic variance can't be explained using genetic features alone. Future work will have to explore how additional features, such as epigenetic features, environmental cues and expression patterns, can be modeled and used in eQTL studies, and how each of these features influences and interacts with other features.

CONEXIC was the first computational model to incorporate DNA features (copy number aberrations) with expression profiles to identify driver genes. CONEXIC, however, is the just the beginning and much more work has to be done in order to effectively identify drivers. The true underlying biology is far more complex, and other genetic aberrations contribute to tumorigenesis. Models that do not incorporate these features are unlikely to find all driver genes, and very likely to report many false-positives. Moreover, expression patterns alone, as demonstrated by COSPER, do not represent all the phenotypes regulated by driver genes. Functional assays, such as siRNA and drug screens, can supplement genomic data and support driver identification.

Explaining the heterogeneous responses to drug treatments is probably the most critical and burning question in the cancer field. My research has just touched the field of drug sensitivity, and

used a very small dataset to do so. Larger datasets with additional data types can greatly enhance our ability to investigate the phenotypic heterogeneity.

One of the greatest challenges I faced while analyzing the gene expression data was the lack of knowledge regarding targets of pathways and transcription factors (TFs). Although a target list for almost every TF is available in at least one study, my data show that targets are context-specific, and genes are regulated by different transcription factors in different cell lines and cell types. A thoroughly collected database aimed at identifying TF targets in different contexts is, in my view, the single most important data that should be collected. Such database will transform the way expression data is analyzed, and will finally allow us to exploit the information hidden in large expression data cohorts.

Conclusions

High throughput data are of great importance to biology research. Sequencing, gene expression, ChIP-seq and similar data types have transformed the way research is performed and significantly enhanced our abilities and knowledge. Statistics and computational modeling provide strong and necessary capabilities to analyze such enormous data sets.

My results in three different projects provide a glimpse into the vast power of computational tools. With GOLPH I showed that non-linear interactions of genetic loci can explain part of the "missing heritability" of GWAS. In CONEXIC we demonstrated that combining data of different types greatly enhances the amount of information one can extract from high-throughput data. With COSPER I showed that a careful examination of transcriptional targets of an oncogenic pathway can help with the identification of resistance mechanisms in cancer.

However, with computational biology, as with any other tool used by biologists, one has to properly select the tool to use, carefully design and collect its input and cautiously interpret its results. However, due to the complexity of computational methods, it is easy to improperly apply them and interpret their results, hindering their great strength and damaging their reputation.

Only by combining knowledge and insights from both computational and experimental biologists we will be able to fully harness the strengths of high-throughput technologies and computational biology.

References

1. Hindorff LA, Sethupathy P, Junkins HA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences of the United States of America 2009;106:9362-7.
2. Lee KW, Woon PS, Teo YY, Sim K. Genome wide association studies (GWAS) and copy number variation (CNV) studies of the major psychoses: what have we learnt? Neuroscience and biobehavioral reviews 2012;36:556-71.
3. Evangelou E, Ioannidis JP. Meta-analysis methods for genome-wide association studies and beyond. Nat Rev Genet 2013;14:379-89.
4. Kato N, Takeuchi F, Tabara Y, et al. Meta-analysis of genome-wide association studies identifies common variants associated with blood pressure variation in east Asians. Nature genetics 2011;43:531-8.
5. Schizophrenia Psychiatric Genome-Wide Association Study C. Genome-wide association study identifies five new schizophrenia loci. Nature genetics 2011;43:969-76.
6. Slatkin M. Epigenetic inheritance and the missing heritability problem. Genetics 2009;182:845-50.
7. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. Nature 2009;461:747-53.
8. Zuk O, Hechter E, Sunyaev SR, Lander ES. The mystery of missing heritability: Genetic interactions create phantom heritability. Proceedings of the National Academy of Sciences of the United States of America 2012;109:1193-8.
9. van Dongen J, Boomsma DI. The evolutionary paradox and the missing heritability of schizophrenia. American journal of medical genetics Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics 2013;162B:122-36.
10. Shilatifard A. Chromatin modifications by methylation and ubiquitination: implications in the regulation of gene expression. Annual review of biochemistry 2006;75:243-69.
11. Barenboim M, Manke T. ChroMoS: an integrated web tool for SNP classification, prioritization and functional interpretation. Bioinformatics 2013;29:2197-8.
12. Akavia UD, Litvin O, Kim J, et al. An integrated approach to uncover drivers of cancer. Cell 2010;143:1005-17.
13. Lee E, Chuang HY, Kim JW, Ideker T, Lee D. Inferring pathway activity toward precise disease classification. PLoS Comput Biol 2008;4:e1000217.
14. Su B, Karin M. Mitogen-activated protein kinase cascades and regulation of gene expression. Current opinion in immunology 1996;8:402-11.
15. Brem RB, Yvert G, Clinton R, Kruglyak L. Genetic dissection of transcriptional regulation in budding yeast. Science 2002;296:752-5.
16. Gerrits A, Li Y, Tesson BM, et al. Expression quantitative trait loci are highly sensitive to cellular differentiation state. PLoS Genet 2009;5:e1000692.
17. Brem RB, Storey JD, Whittle J, Kruglyak L. Genetic interactions between polymorphisms that affect gene expression in yeast. Nature 2005;436:701-3.
18. Cantor RM, Lange K, Sinsheimer JS. Prioritizing GWAS results: A review of statistical methods and recommendations for their application. Am J Hum Genet 2010;86:6-22.
19. Chin L, Garraway LA, Fisher DE. Malignant melanoma: genetics and therapeutics in the genomic era. Genes & development 2006;20:2149-82.
20. Kreimer A, Litvin O, Hao K, Molony C, Pe'er D, Pe'er I. Inference of modules associated to eQTLs. Nucleic Acids Res 2012;40:e98.
21. Davies H, Bignell GR, Cox C, et al. Mutations of the BRAF gene in human cancer. Nature 2002;417:949-54.
22. Latta EK, Tjan S, Parkes RK, O'Malley FP. The role of HER2/neu overexpression/amplification in the progression of ductal carcinoma in situ to invasive carcinoma of the breast. Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc 2002;15:1318-25.
23. Brennan CW, Verhaak RG, McKenna A, et al. The somatic genomic landscape of glioblastoma. Cell 2013;155:462-77.

24. Cancer Genome Atlas Research N. Integrated genomic analyses of ovarian carcinoma. Nature 2011;474:609-15.
25. Beroukhim R, Getz G, Nghiemphu L, et al. Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. Proceedings of the National Academy of Sciences of the United States of America 2007;104:20007-12.
26. Verhaak RG, Hoadley KA, Purdom E, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer cell 2010;17:98-110.
27. Weir BA, Woo MS, Getz G, et al. Characterizing the cancer genome in lung adenocarcinoma. Nature 2007;450:893-8.
28. Flaherty KT, Robert C, Hersey P, et al. Improved survival with MEK inhibition in BRAF-mutated melanoma. The New England journal of medicine 2012;367:107-14.
29. Chapman PB, Hauschild A, Robert C, et al. Improved survival with vemurafenib in melanoma with BRAF V600E mutation. The New England journal of medicine 2011;364:2507-16.
30. Sosman JA, Kim KB, Schuchter L, et al. Survival in BRAF V600-mutant advanced melanoma treated with vemurafenib. The New England journal of medicine 2012;366:707-14.
31. Meloche S, Pouyssegur J. The ERK1/2 mitogen-activated protein kinase pathway as a master regulator of the G1- to S-phase transition. Oncogene 2007;26:3227-39.
32. Dhillon AS, Hagan S, Rath O, Kolch W. MAP kinase signalling pathways in cancer. Oncogene 2007;26:3279-90.
33. Sullivan RJ, Flaherty K. MAP kinase signaling and inhibition in melanoma. Oncogene 2013;32:2373-9.
34. Carriere A, Cargnello M, Julien LA, et al. Oncogenic MAPK signaling stimulates mTORC1 activity by promoting RSK-mediated raptor phosphorylation. Current biology : CB 2008;18:1269-77.
35. Loboda A, Nebozhyn M, Klinghoffer R, et al. A gene expression signature of RAS pathway dependence predicts response to PI3K and RAS pathway inhibitors and expands the population of RAS pathway activated tumors. BMC medical genomics 2010;3:26.
36. Halilovic E, She QB, Ye Q, et al. PIK3CA mutation uncouples tumor growth and cyclin D1 regulation from MEK/ERK and mutant KRAS signaling. Cancer research 2010;70:6804-14.
37. Bloethner S, Chen B, Hemminki K, et al. Effect of common B-RAF and N-RAS mutations on global gene expression in melanoma cell lines. Carcinogenesis 2005;26:1224-32.
38. Roux PP, Blenis J. ERK and p38 MAPK-activated protein kinases: a family of protein kinases with diverse biological functions. Microbiology and molecular biology reviews : MMBR 2004;68:320-44.
39. Dudley DT, Pang L, Decker SJ, Bridges AJ, Saltiel AR. A synthetic inhibitor of the mitogen-activated protein kinase cascade. Proceedings of the National Academy of Sciences of the United States of America 1995;92:7686-9.
40. Tsai J, Lee JT, Wang W, et al. Discovery of a selective inhibitor of oncogenic B-Raf kinase with potent antimelanoma activity. Proceedings of the National Academy of Sciences of the United States of America 2008;105:3041-6.
41. Morris EJ, Jha S, Restaino CR, et al. Discovery of a novel ERK inhibitor with activity in models of acquired resistance to BRAF and MEK inhibitors. Cancer discovery 2013;3:742-50.
42. Xing F, Persaud Y, Pratilas CA, et al. Concurrent loss of the PTEN and RB1 tumor suppressors attenuates RAF dependence in melanomas harboring (V600E)BRAF. Oncogene 2012;31:446-57.
43. Conrad WH, Swift RD, Biechele TL, Kulikauskas RM, Moon RT, Chien AJ. Regulating the response to targeted MEK inhibition in melanoma: enhancing apoptosis in NRAS- and BRAF-mutant melanoma cells with Wnt/beta-catenin activation. Cell Cycle 2012;11:3724-30.
44. Haq R, Yokoyama S, Hawryluk EB, et al. BCL2A1 is a lineage-specific antiapoptotic melanoma oncogene that confers resistance to BRAF inhibition. Proceedings of the National Academy of Sciences of the United States of America 2013;110:4321-6.
45. Barretina J, Caponigro G, Stransky N, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 2012;483:603-7.
46. Garnett MJ, Edelman EJ, Heidorn SJ, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature 2012;483:570-5.

47. Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP. Causal protein-signaling networks derived from multiparameter single-cell data. Science 2005;308:523-9.
48. Siegel R. Cancer Statistics. A Cancer Journal for Clinicians 2013.
49. Clark WH, Jr., Elder DE, Guerry Dt, Epstein MN, Greene MH, Van Horn M. A study of tumor progression: the precursor lesions of superficial spreading and nodular melanoma. Human pathology 1984;15:1147-65.
50. Fearon ER, Hamilton SR, Vogelstein B. Clonal analysis of human colorectal tumors. Science 1987;238:193-7.
51. Berger MF, Hodis E, Heffernan TP, et al. Melanoma genome sequencing reveals frequent PREX2 mutations. Nature 2012;485:502-6.
52. Turajlic S, Furney SJ, Lambros MB, et al. Whole genome sequencing of matched primary and metastatic acral melanomas. Genome Res 2012;22:196-207.
53. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell 2011;144:646-74.
54. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Jr., Kinzler KW. Cancer genome landscapes. Science 2013;339:1546-58.
55. Taparowsky E, Suard Y, Fasano O, Shimizu K, Goldfarb M, Wigler M. Activation of the T24 bladder carcinoma transforming gene is linked to a single amino acid change. Nature 1982;300:762-5.
56. Pollock PM, Harper UL, Hansen KS, et al. High frequency of BRAF mutations in nevi. Nature genetics 2003;33:19-20.
57. Karakas B, Bachman KE, Park BH. Mutation of the PIK3CA oncogene in human cancers. British journal of cancer 2006;94:455-9.
58. Olivier M, Hollstein M, Hainaut P. TP53 mutations in human cancers: origins, consequences, and clinical use. Cold Spring Harbor perspectives in biology 2010;2:a001008.
59. Takahashi T, D'Amico D, Chiba I, Buchhagen DL, Minna JD. Identification of intronic point mutations as an alternative mechanism for p53 inactivation in lung cancer. The Journal of clinical investigation 1990;86:363-9.
60. Spinelli R, Pirola A, Redaelli S, et al. Identification of novel point mutations in splicing sites integrating whole-exome and RNA-seq data in myeloproliferative diseases. Molecular genetics & genomic medicine 2013;1:246-59.
61. Beroukhim R, Mermel CH, Porter D, et al. The landscape of somatic copy-number alteration across human cancers. Nature 2010;463:899-905.
62. Gordon DJ, Resio B, Pellman D. Causes and consequences of aneuploidy in cancer. Nat Rev Genet 2012;13:189-203.
63. Barnard DR, Kalousek DK, Wiersma SR, et al. Morphologic, immunologic, and cytogenetic classification of acute myeloid leukemia and myelodysplastic syndrome in childhood: a report from the Childrens Cancer Group. Leukemia 1996;10:5-12.
64. Bilous M, Morey AL, Armes JE, et al. Assessing HER2 amplification in breast cancer: findings from the Australian In Situ Hybridization Program. Breast cancer research and treatment 2012;134:617-24.
65. Haluska FG, Tsujimoto Y, Croce CM. The t(8;14) chromosome translocation of the Burkitt lymphoma cell line Daudi occurred during immunoglobulin gene rearrangement and involved the heavy chain diversity region. Proceedings of the National Academy of Sciences of the United States of America 1987;84:6835-9.
66. Honeyman JN, Simon EP, Robine N, et al. Detection of a recurrent DNAJB1-PRKACA chimeric transcript in fibrolamellar hepatocellular carcinoma. Science 2014;343:1010-4.
67. Rowley JD. Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. Nature 1973;243:290-3.
68. Biswas S, Trobridge P, Romero-Gallo J, et al. Mutational inactivation of TGFBR2 in microsatellite unstable colon cancer arises from the cooperation of genomic instability and the clonal outgrowth of transforming growth factor beta resistant cells. Genes, chromosomes & cancer 2008;47:95-106.
69. Hartwell L. Defects in a cell cycle checkpoint may be responsible for the genomic instability of cancer cells. Cell 1992;71:543-6.

70. Loeb LA. Mutator phenotype may be required for multistage carcinogenesis. Cancer research 1991;51:3075-9.
71. Tomlinson IP, Novelli MR, Bodmer WF. The mutation rate and cancer. Proceedings of the National Academy of Sciences of the United States of America 1996;93:14800-3.
72. Peltomaki P, de la Chapelle A. Mutations predisposing to hereditary nonpolyposis colorectal cancer. Advances in cancer research 1997;71:93-119.
73. Papadopoulos N, Lindblom A. Molecular basis of HNPCC: mutations of MMR genes. Human mutation 1997;10:89-99.
74. Campeau PM, Foulkes WD, Tischkowitz MD. Hereditary breast cancer: new genetic developments, new therapeutic avenues. Human genetics 2008;124:31-42.
75. Easton DF. How many more breast cancer predisposition genes are there? Breast cancer research : BCR 1999;1:14-7.
76. Stewart SA, Weinberg RA. Telomeres: cancer to human aging. Annual review of cell and developmental biology 2006;22:531-57.
77. Stephens P, Edkins S, Davies H, et al. A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer. Nature genetics 2005;37:590-2.
78. McCready S, Carr AM, Lehmann AR. Repair of cyclobutane pyrimidine dimers and 6-4 photoproducts in the fission yeast Schizosaccharomyces pombe. Mol Microbiol 1993;10:885-90.
79. Liu XY, Zhu MX, Xie JP. Mutagenicity of acrolein and acrolein-induced DNA adducts. Toxicology mechanisms and methods 2010;20:36-44.
80. Nouspikel T. DNA repair in mammalian cells : Nucleotide excision repair: variations on versatility. Cellular and molecular life sciences : CMLS 2009;66:994-1009.
81. Friedberg EC. How nucleotide excision repair protects against cancer. Nature reviews Cancer 2001;1:22-33.
82. Kunkel TA, Bebenek K. DNA replication fidelity. Annual review of biochemistry 2000;69:497-529.
83. Preston BD, Albertson TM, Herr AJ. DNA replication fidelity and cancer. Seminars in cancer biology 2010;20:281-93.
84. Imai K, Yamamoto H. Carcinogenesis and microsatellite instability: the interrelationship between genetics and epigenetics. Carcinogenesis 2008;29:673-80.
85. Gryfe R, Kim H, Hsieh ET, et al. Tumor microsatellite instability and clinical outcome in young patients with colorectal cancer. The New England journal of medicine 2000;342:69-77.
86. F M, B J, F M. Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer. 2014.
87. Vogelstein B, Fearon ER, Kern SE, et al. Allelotype of colorectal carcinomas. Science 1989;244:207-11.
88. Cahill DP, Lengauer C, Yu J, et al. Mutations of mitotic checkpoint genes in human cancers. Nature 1998;392:300-3.
89. Albertson DG. Gene amplification in cancer. Trends in genetics : TIG 2006;22:447-55.
90. Paulson TG, Almasan A, Brody LL, Wahl GM. Gene amplification in a p53-deficient cell line requires cell cycle progression under conditions that generate DNA breakage. Mol Cell Biol 1998;18:3089-100.
91. Comings DE. A general theory of carcinogenesis. Proceedings of the National Academy of Sciences of the United States of America 1973;70:3324-8.
92. Knudson AG, Jr. Mutation and cancer: statistical study of retinoblastoma. Proceedings of the National Academy of Sciences of the United States of America 1971;68:820-3.
93. Yunis JJ, Ramsay N. Retinoblastoma and subband deletion of chromosome 13. American journal of diseases of children 1978;132:161-3.
94. Malumbres M, Barbacid M. RAS oncogenes: the first 30 years. Nature reviews Cancer 2003;3:459-65.
95. Cleaver JE. Cancer in xeroderma pigmentosum and related disorders of DNA repair. Nature reviews Cancer 2005;5:564-73.
96. Beerenwinkel N, Antal T, Dingli D, et al. Genetic progression and the waiting time to cancer. PLoS Comput Biol 2007;3:e225.

97. Miller DG. On the nature of susceptibility to cancer. The presidential address. Cancer 1980;46:1307-18.
98. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature 2009;458:719-24.
99. Samuels Y, Wang Z, Bardelli A, et al. High frequency of mutations of the PIK3CA gene in human cancers. Science 2004;304:554.
100. Dhomen N, Marais R. BRAF signaling and targeted therapies in melanoma. Hematology/oncology clinics of North America 2009;23:529-45, ix.
101. Lander ES. Initial impact of the sequencing of the human genome. Nature 2011;470:187-97.
102. Parsons DW, Jones S, Zhang X, et al. An integrated genomic analysis of human glioblastoma multiforme. Science 2008;321:1807-12.
103. Garraway LA, Widlund HR, Rubin MA, et al. Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. Nature 2005;436:117-22.
104. Sanchez-Garcia F, Akavia UD, Mozes E, Pe'er D. JISTIC: identification of significant targets in cancer. BMC Bioinformatics 2010;11:189.
105. Cancer Genome Atlas N. Comprehensive molecular portraits of human breast tumours. Nature 2012;490:61-70.
106. Noy NF, Shah NH, Whetzel PL, et al. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Res 2009;37:W170-3.
107. Bokemeyer A, Eckert C, Meyr F, et al. Copy number genome alterations are associated with treatment response and outcome in relapsed childhood ETV6/RUNX1-positive acute lymphoblastic leukemia. Haematologica 2013.
108. Engler DA, Gupta S, Growdon WB, et al. Genome wide DNA copy number analysis of serous type ovarian carcinomas identifies genetic markers predictive of clinical outcome. PloS one 2012;7:e30996.
109. Huang YT, Heist RS, Chirieac LR, et al. Genome-wide analysis of survival in early-stage non-small-cell lung cancer. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 2009;27:2660-7.
110. Xie T, G DA, Lamb JR, et al. A comprehensive characterization of genome-wide copy number aberrations in colorectal cancer reveals novel oncogenes and patterns of alterations. PloS one 2012;7:e42001.
111. Dat le T, Matsuo T, Yoshimaru T, et al. Identification of genes potentially involved in bone metastasis by genome-wide gene expression profile analysis of non-small cell lung cancer in mice. International journal of oncology 2012;40:1455-69.
112. Hicks C, Kumar R, Pannuti A, Miele L. Integrative Analysis of Response to Tamoxifen Treatment in ER-Positive Breast Cancer Using GWAS Information and Transcription Profiling. Breast cancer : basic and clinical research 2012;6:47-66.
113. Weichselbaum RR, Ishwaran H, Yoon T, et al. An interferon-related gene signature for DNA damage resistance is a predictive marker for chemotherapy and radiation for breast cancer. Proceedings of the National Academy of Sciences of the United States of America 2008;105:18490-5.
114. Prahallad A, Sun C, Huang S, et al. Unresponsiveness of colon cancer to BRAF(V600E) inhibition through feedback activation of EGFR. Nature 2012;483:100-3.
115. Mendes-Pereira AM, Sims D, Dexter T, et al. Genome-wide functional screen identifies a compendium of genes affecting sensitivity to tamoxifen. Proceedings of the National Academy of Sciences of the United States of America 2012;109:2730-5.
116. Cheung HW, Cowley GS, Weir BA, et al. Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. Proceedings of the National Academy of Sciences of the United States of America 2011;108:12372-7.
117. Madhunapantula SV, Robertson GP. The PTEN-AKT3 signaling cascade as a therapeutic target in melanoma. Pigment cell & melanoma research 2009;22:400-19.
118. Carvajal RD, Antonescu CR, Wolchok JD, et al. KIT as a therapeutic target in metastatic melanoma. JAMA 2011;305:2327-34.
119. Levy C, Khaled M, Fisher DE. MITF: master regulator of melanocyte development and melanoma oncogene. Trends in molecular medicine 2006;12:406-14.

120. Javelaud D, Alexaki VI, Mauviel A. Transforming growth factor-beta in cutaneous melanoma. Pigment cell & melanoma research 2008;21:123-32.
121. Rubinfeld B, Robbins P, El-Gamil M, Albert I, Porfiri E, Polakis P. Stabilization of beta-catenin by genetic defects in melanoma cell lines. Science 1997;275:1790-2.
122. Monzon J, Liu L, Brill H, et al. CDKN2A mutations in multiple primary melanomas. The New England journal of medicine 1998;338:879-87.
123. Pansky A, Hildebrand P, Fasler-Kan E, et al. Defective Jak-STAT signal transduction pathway in melanoma cells resistant to growth inhibition by interferon-alpha. International journal of cancer Journal international du cancer 2000;85:720-5.
124. Lopez-Bergami P, Fitchman B, Ronai Z. Understanding signaling cascades in melanoma. Photochemistry and photobiology 2008;84:289-306.
125. Omholt K, Platz A, Kanter L, Ringborg U, Hansson J. NRAS and BRAF mutations arise early during melanoma pathogenesis and are preserved throughout tumor progression. Clinical cancer research : an official journal of the American Association for Cancer Research 2003;9:6483-8.
126. Solit DB, Garraway LA, Pratilas CA, et al. BRAF mutation predicts sensitivity to MEK inhibition. Nature 2006;439:358-62.
127. Flaherty KT, Puzanov I, Kim KB, et al. Inhibition of mutated, activated BRAF in metastatic melanoma. The New England journal of medicine 2010;363:809-19.
128. Kortylewski M, Heinrich PC, Kauffmann ME, Bohm M, MacKiewicz A, Behrmann I. Mitogen-activated protein kinases control p27/Kip1 expression and growth of human melanoma cells. The Biochemical journal 2001;357:297-303.
129. Lefevre G, Calipel A, Mouriaux F, Hecquet C, Malecaze F, Mascarelli F. Opposite long-term regulation of c-Myc and p27Kip1 through overactivation of Raf-1 and the MEK/ERK module in proliferating human choroidal melanoma cells. Oncogene 2003;22:8813-22.
130. Pratilas CA, Taylor BS, Ye Q, et al. (V600E)BRAF is associated with disabled feedback inhibition of RAF-MEK signaling and elevated transcriptional output of the pathway. Proceedings of the National Academy of Sciences of the United States of America 2009;106:4519-24.
131. Cartlidge RA, Thomas GR, Cagnol S, et al. Oncogenic BRAF(V600E) inhibits BIM expression to promote melanoma cell survival. Pigment cell & melanoma research 2008;21:534-44.
132. Romeo Y, Moreau J, Zindy PJ, et al. RSK regulates activated BRAF signalling to mTORC1 and promotes melanoma growth. Oncogene 2013;32:2917-26.
133. Bollag G, Hirth P, Tsai J, et al. Clinical efficacy of a RAF inhibitor needs broad target blockade in BRAF-mutant melanoma. Nature 2010;467:596-9.
134. Flaherty KT, Infante JR, Daud A, et al. Combined BRAF and MEK inhibition in melanoma with BRAF V600 mutations. The New England journal of medicine 2012;367:1694-703.
135. Gray-Schopfer V, Wellbrock C, Marais R. Melanoma biology and new targeted therapy. Nature 2007;445:851-7.
136. Salti GI, Manougian T, Farolan M, Shilkaitis A, Majumdar D, Das Gupta TK. Micropthalmia transcription factor: a new prognostic marker in intermediate-thickness cutaneous malignant melanoma. Cancer research 2000;60:5012-6.
137. Johannessen CM, Johnson LA, Piccioni F, et al. A melanocyte lineage program confers resistance to MAP kinase pathway inhibition. Nature 2013;504:138-42.
138. Kennedy C, ter Huurne J, Berkhout M, et al. Melanocortin 1 receptor (MC1R) gene variants are associated with an increased risk for cutaneous melanoma which is largely independent of skin type and hair color. The Journal of investigative dermatology 2001;117:294-300.
139. Frandberg PA, Doufexis M, Kapas S, Chhajlani V. Human pigmentation phenotype: a point mutation generates nonfunctional MSH receptor. Biochemical and biophysical research communications 1998;245:490-2.
140. Price ER, Ding HF, Badalian T, et al. Lineage-specific signaling in melanocytes. C-kit stimulation recruits p300/CBP to microphthalmia. The Journal of biological chemistry 1998;273:17983-6.
141. Miller AJ, Levy C, Davis IJ, Razin E, Fisher DE. Sumoylation of MITF and its related family members TFE3 and TFEB. The Journal of biological chemistry 2005;280:146-55.

142. Wu M, Hemesath TJ, Takemoto CM, et al. c-Kit triggers dual phosphorylations, which couple activation and degradation of the essential melanocyte factor Mi. Genes & development 2000;14:301-12.
143. Aitken J, Welch J, Duffy D, et al. CDKN2A variants in a population-based sample of Queensland families with melanoma. J Natl Cancer Inst 1999;91:446-52.
144. Castellano M, Pollock PM, Walters MK, et al. CDKN2A/p16 is inactivated in most melanoma cell lines. Cancer research 1997;57:4868-75.
145. Flores JF, Walker GJ, Glendening JM, et al. Loss of the p16INK4a and p15INK4b genes, as well as neighboring 9p21 markers, in sporadic melanoma. Cancer research 1996;56:5023-32.
146. Sauter ER, Yeo UC, von Stemm A, et al. Cyclin D1 is a candidate oncogene in cutaneous melanoma. Cancer research 2002;62:3200-6.
147. Li W, Sanki A, Karim RZ, et al. The role of cell cycle regulatory proteins in the pathogenesis of melanoma. Pathology 2006;38:287-301.
148. Musgrove EA, Caldon CE, Barraclough J, Stone A, Sutherland RL. Cyclin D as a therapeutic target in cancer. Nature reviews Cancer 2011;11:558-72.
149. Wu H, Goel V, Haluska FG. PTEN signaling pathways in melanoma. Oncogene 2003;22:3113-22.
150. Guldberg P, thor Straten P, Birck A, Ahrenkiel V, Kirkin AF, Zeuthen J. Disruption of the MMAC1/PTEN gene by deletion or mutation is a frequent event in malignant melanoma. Cancer research 1997;57:3660-3.
151. Stahl JM, Sharma A, Cheung M, et al. Deregulated Akt3 activity promotes development of malignant melanoma. Cancer research 2004;64:7002-10.
152. Engelman JA. Targeting PI3K signalling in cancer: opportunities, challenges and limitations. Nature reviews Cancer 2009;9:550-62.
153. Paraiso KH, Xiang Y, Rebecca VW, et al. PTEN loss confers BRAF inhibitor resistance to melanoma cells through the suppression of BIM expression. Cancer research 2011;71:2750-60.
154. Kwong LN, Davies MA. Targeted therapy for melanoma: rational combinatorial approaches. Oncogene 2014;33:1-9.
155. Gopal YN, Deng W, Woodman SE, et al. Basal and treatment-induced activation of AKT mediates resistance to cell death by AZD6244 (ARRY-142886) in Braf-mutant human cutaneous melanoma cells. Cancer research 2010;70:8736-47.
156. Shi H, Kong X, Ribas A, Lo RS. Combinatorial treatments that overcome PDGFRbeta-driven resistance of melanoma cells to V600EB-RAF inhibition. Cancer research 2011;71:5067-74.
157. Dong Y, Richards JA, Gupta R, et al. PTEN functions as a melanoma tumor suppressor by promoting host immune response. Oncogene 2013.
158. Litvin O, Causton HC, Chen BJ, Pe'er D. Modularity and interactions in the genetics of gene expression. Proceedings of the National Academy of Sciences of the United States of America 2009;106:6441-6.
159. Wellcome Trust Case Control C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007;447:661-78.
160. Cheung VG, Conlin LK, Weber TM, et al. Natural variation in human gene expression assessed in lymphoblastoid cells. Nature genetics 2003;33:422-5.
161. Schadt EE, Monks SA, Drake TA, et al. Genetics of gene expression surveyed in maize, mouse and man. Nature 2003;422:297-302.
162. Yvert G, Brem RB, Whittle J, et al. *Trans*-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. Nature genetics 2003;35:57-64.
163. Brem RB, Kruglyak L. The landscape of genetic complexity across 5,700 gene expression traits in yeast. Proceedings of the National Academy of Sciences of the United States of America 2005;102:1572-7.
164. Smith EN, Kruglyak L. Gene-environment interaction in yeast gene expression. PLoS Biol 2008;6:e83.
165. Storey JD, Akey JM, Kruglyak L. Multiple locus linkage analysis of genomewide expression in yeast. PLoS Biol 2005;3:e267.
166. Zhu J, Zhang B, Smith EN, et al. Integrating large-scale functional genomic dadta to dissect the complexity of yeast regulatory networks. Nat Genet 2008;40:854-61.

167. Perlstein EO, Ruderfer DM, Roberts DC, Schreiber SL, Kruglyak L. Genetic basis of individual differences in the response to small-molecule drugs in yeast. Nature genetics 2007;39:496-502.
168. Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. Nature 1999;402:C47-52.
169. Segal E, Shapira M, Regev A, et al. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nature genetics 2003;34:166-76.
170. Hughes JD, Estep PW, Tavazoie S, Church GM. Computational Identification of *Cis*-regulatory Elements Associated with Groups of Functionally Related Genes in *Saccharomyces cerevisiae*. J Mol Biol 2000;296:1205-14.
171. Kwast KE, Burke PV, Poyton RO. Oxygen sensing and the transcriptional regulation of oxygen-responsive genes in yeast. J Exp Biol 1998;201:1177-95.
172. Dibrov E, Fu S, Lemire BD. The Saccharomyces cerevisiae TCM62 gene encodes a chaperone necessary for the assembly of the mitochondrial succinate dehydrogenase (complex II). The Journal of biological chemistry 1998;273:32042-8.
173. MacIsaac KD, Wang T, Gordon DB, Gifford DK, Stormo GD, Fraenkel E. An improved map of conserved regulatory sites for Saccharomyces cerevisiae. BMC Bioinformatics 2006;7:113.
174. Lee SI, Pe'er D, Dudley AM, Church GM, Koller D. Identifying regulatory mechanisms using individual variation reveals key role for chromatin modification. Proceedings of the National Academy of Sciences of the United States of America 2006;103:14062-7.
175. Hu Z, Killion PJ, Iyer VR. Genetic reconstruction of a functional transcriptional regulatory network. Nature genetics 2007;39:683-7.
176. Roberts CJ, Nelson B, Marton MJ, et al. Signaling and Circuitry of Multiple MAPK Pathways Revealed by a Matrix of Global Gene Expression Profiles. Science 2000;287:873-80.
177. Lee TI, Causton HC, Holstege FCP, et al. Redundant Roles for the TFIID and SAGA Complexes in Global Transcription. Nature 2000;405:701-4.
178. Lee TI, Rinaldi NJ, Robert F, et al. Transcriptional Regulatory Networks in Saccharomyces cerevisiae. Science 2002;298:799-804.
179. Lander E, Kruglyak L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nature genetics 1995;11:241-7.
180. Welch BL. The significance of the difference between two means when the population variances are unequal. Biometrika 1938;29:350-62.
181. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences of the United States of America 2003;100:9440-5.
182. Harbison CT, Gordon DB, Lee TI, et al. Transcriptional regulatory code of a eukaryotic genome. Nature 2004;431:99-104.
183. Lieb JD, Liu X, Botstein D, Brown PO. Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. Nature genetics 2001;28:327-34.
184. Gelbart ME, Bachman N, Delrow J, Boeke JD, Tsukiyama T. Genome-wide identification of Isw2 chromatin-remodeling targets by localization of a catalytically inactive mutant. Genes & development 2005;19:942-54.
185. Hughes TR, Marton MJ, Jones AR, et al. Functional Discovery via a Compendium of Expression Profiles. Cell 2000;102:109-26.
186. Chen BJ, Causton HC, Mancenido D, Goddard NL, Perlstein EO, Pe'er D. Harnessing gene expression to identify the genetic basis of drug resistance. Mol Syst Biol 2009;5:310.
187. Lin WM, Baker AC, Beroukhim R, et al. Modeling genomic diversity and tumor dependency in malignant melanoma. Cancer research 2008;68:664-73.
188. Beroukhim R, Brunet JP, Di Napoli A, et al. Patterns of gene expression and copy-number alterations in von-hippel lindau disease-associated and sporadic clear cell carcinoma of the kidney. Cancer research 2009;69:4674-81.
189. Golub TR, Slonim D, Tamayo P, et al. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science 1999;286:531-7.
190. Segal E, Friedman N, Koller D, Regev A. A module map showing conditional activity of expression modules in cancer. Nature genetics 2004;36:1090-8.

191. Huh SJ, Chen YL, Friedman SL, et al. KLF6 Gene and early melanoma development in a collagen I-rich extracellular environment. J Natl Cancer Inst 2010;102:1131-47.
192. Adler AS, Lin M, Horlings H, Nuyten DS, van de Vijver MJ, Chang HY. Genetic regulators of large-scale transcriptional signatures in cancer. Nature genetics 2006;38:421-30.
193. Hoek KS, Schlegel NC, Eichhoff OM, et al. Novel MITF targets identified using a two-step DNA microarray strategy. Pigment cell & melanoma research 2008;21:665-76.
194. Witten DM, Tibshirani R, Hastie T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics 2009;10:515-34.
195. Raychaudhuri S, Plenge RM, Rossin EJ, et al. Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. PLoS Genet 2009;5:e1000534.
196. Cerami E, Demir E, Schultz N, Taylor BS, Sander C. Automated network analysis identifies core pathways in glioblastoma. PloS one 2010;5:e8918.
197. Wang K, Saito M, Bisikirska BC, et al. Genome-wide identification of post-translational modulators of transcription factor activity in human B cells. Nat Biotechnol 2009;27:829-39.
198. Hoek KS, Schlegel NC, Brafford P, et al. Metastatic potential of melanomas defined by specific gene expression profiles with no BRAF signature. Pigment cell research / sponsored by the European Society for Pigment Cell Research and the International Pigment Cell Society 2006;19:290-302.
199. Johansson P, Pavey S, Hayward N. Confirmation of a BRAF mutation-associated gene expression signature in melanoma. Pigment cell research / sponsored by the European Society for Pigment Cell Research and the International Pigment Cell Society 2007;20:216-21.
200. Vallabhapurapu S, Matsuzawa A, Zhang W, et al. Nonredundant and complementary functions of TRAF2 and TRAF3 in a ubiquitination cascade that activates NIK-dependent alternative NF-kappaB signaling. Nat Immunol 2008;9:1364-70.
201. Steingrimsson E, Copeland NG, Jenkins NA. Melanocytes and the microphthalmia transcription factor network. Annual review of genetics 2004;38:365-411.
202. Du J, Widlund HR, Horstmann MA, et al. Critical role of CDK2 for melanoma growth linked to its melanocyte-specific transcriptional regulation by MITF. Cancer cell 2004;6:565-76.
203. Itoh T, Satoh M, Kanno E, Fukuda M. Screening for target Rabs of TBC (Tre-2/Bub2/Cdc16) domain-containing proteins based on their Rab-binding activity. Genes Cells 2006;11:1023-37.
204. Satijn DP, Olson DJ, van der Vlag J, et al. Interference with the expression of a novel human polycomb protein, hPc2, results in cellular transformation and apoptosis. Mol Cell Biol 1997;17:6076-86.
205. Jordens I, Westbroek W, Marsman M, et al. Rab7 and Rab27a control two motor protein activities involved in melanosomal transport. Pigment cell research / sponsored by the European Society for Pigment Cell Research and the International Pigment Cell Society 2006;19:412-23.
206. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America 2005;102:15545-50.
207. Kent WJ, Sugnet CW, Furey TS, et al. The human genome browser at UCSC. Genome Res 2002;12:996-1006.
208. Salton G. Automatic text processing : the transformation, analysis, and retrieval of information by computer. Reading, Mass.: Addison-Wesley; 1988.
209. Wellbrock C, Rana S, Paterson H, Pickersgill H, Brummelkamp T, Marais R. Oncogenic BRAF regulates melanoma proliferation through the lineage specific factor MITF. PloS one 2008;3:e2734.
210. Turner N, Lambros MB, Horlings HM, et al. Integrative molecular profiling of triple negative breast cancers identifies amplicon drivers and potential therapeutic targets. Oncogene 2010;29:2013-23.
211. Chiaverini C, Beuret L, Flori E, et al. Microphthalmia-associated transcription factor regulates RAB27A gene expression and controls melanosome transport. The Journal of biological chemistry 2008;283:12635-42.
212. Scott KL, Kabbarah O, Liang MC, et al. GOLPH3 modulates mTOR signalling and rapamycin sensitivity in cancer. Nature 2009;459:1085-90.

213. Piccart-Gebhart MJ, Procter M, Leyland-Jones B, et al. Trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer. The New England journal of medicine 2005;353:1659-72.
214. Joseph EW, Pratilas CA, Poulikakos PI, et al. The RAF inhibitor PLX4032 inhibits ERK signaling and tumor cell proliferation in a V600E BRAF-selective manner. Proceedings of the National Academy of Sciences of the United States of America 2010;107:14903-8.
215. Slamon DJ, Leyland-Jones B, Shak S, et al. Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. The New England journal of medicine 2001;344:783-92.
216. Niepel M, Hafner M, Pace EA, et al. Profiles of Basal and stimulated receptor signaling networks predict drug response in breast cancer lines. Science signaling 2013;6:ra84.
217. Hodis E, Watson IR, Kryukov GV, et al. A landscape of driver mutations in melanoma. Cell 2012;150:251-63.
218. Bracher A, Cardona AS, Tauber S, et al. Epidermal growth factor facilitates melanoma lymph node metastasis by influencing tumor lymphangiogenesis. The Journal of investigative dermatology 2013;133:230-8.
219. Hodi FS, O'Day SJ, McDermott DF, et al. Improved survival with ipilimumab in patients with metastatic melanoma. The New England journal of medicine 2010;363:711-23.
220. Molina DM, Grewal S, Bardwell L. Characterization of an ERK-binding domain in microphthalmia-associated transcription factor and differential inhibition of ERK2-mediated substrate phosphorylation. The Journal of biological chemistry 2005;280:42051-60.
221. Goding CR. Mitf from neural crest to melanoma: signal transduction and transcription in the melanocyte lineage. Genes & development 2000;14:1712-28.
222. Zhang J, Xiao Z, Lai D, et al. miR-21, miR-17 and miR-19a induced by phosphatase of regenerating liver-3 promote the proliferation and metastasis of colon cancer. British journal of cancer 2012;107:352-9.
223. Dai B, Meng J, Peyton M, et al. STAT3 mediates resistance to MEK inhibitor through microRNA miR-17. Cancer research 2011;71:3658-68.
224. Gao K, Dai DL, Martinka M, Li G. Prognostic significance of nuclear factor-kappaB p105/p50 in human melanoma and its role in cell migration. Cancer research 2006;66:8382-8.
225. Platanias LC. Mechanisms of type-I- and type-II-interferon-mediated signalling. Nature reviews Immunology 2005;5:375-86.
226. Jackson DP, Watling D, Rogers NC, et al. The JAK/STAT pathway is not sufficient to sustain the antiproliferative response in an interferon-resistant human melanoma cell line. Melanoma research 2003;13:219-29.
227. Leaman DW, Chawla-Sarkar M, Jacobs B, et al. Novel growth and death related interferon-stimulated genes (ISGs) in melanoma: greater potency of IFN-beta compared with IFN-alpha2. Journal of interferon & cytokine research : the official journal of the International Society for Interferon and Cytokine Research 2003;23:745-56.
228. Li X, Leung S, Qureshi S, Darnell JE, Jr., Stark GR. Formation of STAT1-STAT2 heterodimers and their role in the activation of IRF-1 gene transcription by interferon-alpha. The Journal of biological chemistry 1996;271:5790-4.
229. Bratton SB, Salvesen GS. Regulation of the Apaf-1-caspase-9 apoptosome. Journal of cell science 2010;123:3209-14.
230. Reed JA, Loganzo F, Jr., Shea CR, et al. Loss of expression of the p16/cyclin-dependent kinase inhibitor 2 tumor suppressor gene in melanocytic lesions correlates with invasive stage of tumor progression. Cancer research 1995;55:2713-8.
231. Naylor MF, Brown S, Quinlan C, Pitha JV, Evertt MA. 9p21 deletions in primary melanoma. Dermatology online journal 1997;3:1.
232. DeGroot MH. Optimal statistical decisions. Wiley classics library ed. Hoboken, N.J.: Wiley-Interscience; 2004.
233. Sanceau J, Hiscott J, Delattre O, Wietzerbin J. IFN-beta induces serine phosphorylation of Stat-1 in Ewing's sarcoma cells and mediates apoptosis via induction of IRF-1 and activation of caspase-7. Oncogene 2000;19:3372-83.

234. Corcoran RB, Ebi H, Turke AB, et al. EGFR-mediated re-activation of MAPK signaling contributes to insensitivity of BRAF mutant colorectal cancers to RAF inhibition with vemurafenib. Cancer discovery 2012;2:227-35.
235. Vanlandingham JW, Tassabehji NM, Somers RC, Levenson CW. Expression profiling of p53-target genes in copper-mediated neuronal apoptosis. Neuromolecular medicine 2005;7:311-24.
236. Burns TF, El-Deiry WS. Microarray analysis of p53 target gene expression patterns in the spleen and thymus in response to ionizing radiation. Cancer biology & therapy 2003;2:431-43.